Títol: Extracció intel·ligent del contingut de documents semiestructurats i desestructurats mitjançant tècniques avançades de Captura Digital

Alumne: David Durich Mas

Ponent: Pere Marés Martí

Departament: ESAII

Data: 22 de Gener de 2013

DADES DEL PROJECTE

Títol del Projecte: Extracció intel·ligent del contingut de documents semiestructurats i desestructurats mitjançant tècniques avançades de Captura Digital

Nom de l'estudiant: David Durich Mas
Titulació: Enginyeria en Informàtica
Crèdits: 37,5
Director: Oscar Saro Arroyo
Institució del director: Technology for Business Solutions (TBS)
Ponent: Pere Marés Martí
Departament: ESAII

MEMBRES DEL TRIBUNAL *(nom i signatura)*

President: Juan Climent Vilaró

Vocal: Robert Lukas Mario Nieuwenhuis

Secretari: Pere Marés Martí

QUALIFICACIÓ

Qualificació numèrica:
Qualificació descriptiva:

Data: 22 de Gener de 2013

Índex

1. Introducció

Avui en dia encara s'utilitzen en el món empresarial molts tipus de documents que només s'introdueixen a fluxos d'informació digital de manera manual. Documents de contingut semiestructurat (factures, comandes, albarans…) o desestructurat (documents tècnics, manuals, currículums, notícies…). La manca d'estructura d'aquest tipus de documents dificulta molt l'ús de tècniques d'extracció tradicionals, que en la seva majoria són posicionals.

El present Projecte de Final de Carrera proposa la utilització de tècniques de captura digital que extreguin automàticament el contingut d'aquest tipus de documents basades en cerques sintàctiques i en regles de negoci. Concretament, es proposa una solució per automatitzar la captura i gestió de les factures de proveïdors en un entorn empresarial. Actualment, en aquests tipus de processos documentals encara s'usa majoritàriament el paper i, per tant, una solució com la que es pretén desenvolupar pot agilitzar i abaratir de manera significativa el procés d'administració d'aquests documents.

Els processos que es construiran constaran de varies etapes on el que s'intentarà és optimitzar la gestió documental en matèria de captura digital dins de l'entorn empresarial. En la primera etapa es compararan i es decidiran les diferents tècniques d'extracció digital i de reconeixement de caràcters a utilitzar per tal d'aconseguir una extracció de dades òptima. Posteriorment, es passarà a generar totes les regles lògiques que permetran localitzar, reconèixer i classificar les dades que volem extreure. Com que es tracta de documents semiestructurats, aquestes dues primeres etapes s'hauran de dur a terme a partir d'un conjunt de mostres d'aquests documents que sigui suficientment gran com per poder treballar i configurar els formats o estructures més habituals en els que s'utilitzen i així poder obtenir la solució més acurada possible. Un cop disposem d'aquestes dades, la següent etapa serà la de validar-les automàticament amb regles financeres i amb bases de dades corporatives. Posteriorment, s'habilitarà una interfície amb la que l'usuari podrà revisar i modificar les dades si ho creu necessari. Finalment, es procedirà a emmagatzemar-les i integrar-

les amb un gestor documental, a partir del qual podrem exportar els valors finals al ERP (*Enterprise resource planning*) corporatiu.

2. Objectius

L'objectiu principal del PFC és aconseguir una solució que ens permeti la introducció i gestió de les dades contingudes en documents professionals semiestructurats als sistemes d'informació empresarials habituals mitjançant tècniques avançades de captura digital.

El projecte es desenvolupa en el marc empresarial de la Corporació Albatros, la qual està formada per varies societats que mouen un gran volum de facturació a proveïdors. Totes aquestes factures es reben en paper imprès i són un clar exemple de document semiestructurat. Degut al gran volum d'aquests documents el temps i el cost que s'inverteix a introduir manualment aquestes dades a l'ERP de la corporació són bastant significatius. A més, s'ha expressat la voluntat de minimitzar la duplicitat de factures i els errors numèrics a l'hora de tractar la introducció de les dades. Degut a aquestes circumstàncies es requereix una solució complexa per poder automatitzar el procés d'introducció i validació d'aquestes dades.

Aquesta solució es desenvolupa dins l'entorn de TBS, empresa especialitzada en solucions tecnològiques documentals.

A continuació s'exposen els objectius que es pretenen assolir amb la realització del present projecte:

- Estudi i justificació de les tècniques de reconeixement digital i del software a utilitzar.
- Extracció del contingut dels documents escanejats amb tècniques basades en cerques sintàctiques i regles de localització.
- Desenvolupament de la solució per validar el format de les dades resultants de la extracció.
- Validació dels resultats en conjunt mitjançant regles de negoci i dades mestres contingudes en bases de dades.

- Disseny i desenvolupament del formulari de validació per l'usuari.
- Transferència de les dades a bases de dades i sistemes de negoci (Gestors documentals i ERPs).
- Integració de la digitalització certificada al sistema.
- Anàlisi i valoració de resultats.

3. Introducció a la captura digital

A continuació, es fa una breu descripció de l'escenari en el que ens trobem en matèria de captura digital documental i es defineixen els principals components: l'OCR i la gestió documental.

3.1 Escenari i evolució

Al llarg de la historia i des de la invenció de la impremta tota la gestió de la informació ja sigui d'entitats públiques o privades ha sigut mitjançant el document escrit o imprès. L'acumulació, catalogació i conservació de tots aquests documents també han tingut una particular importància per les cultures a tot el món. Abans de Gütenberg[i] fer una còpia d'una obra escrita requeria la pacient labor d'un o varis copistes que al cap de mesos o anys de feina aconseguien reproduir-la. Amb la impremta es van abaratir molt els costos augmentant així les possibilitats, no només de tenir múltiples copies d'un mateix escrit, sinó de poder entendre la gestió de la informació d'una altra manera, ja que es podia agilitzar i tenir constància de tot tipus de transaccions i processos que la societat anava requerint. Actualment, en un món al que es sol anomenar digital, continuem depenent del document en paper per a la gestió i emmagatzemen de molts tipus d'informació.

Amb l'aparició de les tecnologies de digitalització hem donat un pas més, ja que els avantatges de poder digitalitzar tots aquests documents són enormes. La primera i més evident respon a la necessitat de conservar còpies dels documents en arxius. Aquests, davant de les fotocòpies en paper i en microfilm presenten avantatges tan significatives com la rapidesa en la recuperació de la copia desitjada, recuperació que dependrà de les metadades que s'associïn a la imatge en qüestió.

Però el concepte de digitalització de la informació va més enllà de fer simples còpies de documents en paper per convertir-los i emmagatzemar-los en imatges. Perquè aquesta digitalització pugui ser completa hem de poder tractar la informació que conté i això ho podem fer de manera automàtica amb els sistemes de Reconeixement Òptic de Caràcters (OCR) mitjançant els quals es poden extreure els caràcters d'una imatge digital i indexar-los posteriorment de la manera desitjada.

Per una altra banda, la importància que l'entorn digital ha adquirit en els últims anys, impulsat per la legislació desenvolupada en vers a ell, proporciona als documents digitals les garanties bàsiques de supervivència, manteniment tecnològic i innovació. Un altre avantatge que ofereix aquest format és la possibilitat que la còpia digital sigui firmada electrònicament, originant així l'autenticació d'aquesta i podent entrar a formar part de la producció documental derivada de l'administració electrònica corresponent sense necessitat de conservar la còpia escrita.

3.2 OCR (Optical Character Recognition)

OCR és la digitalització del text escrit o imprès contingut en una imatge escanejada. En les aplicacions software OCR s'utilitzen tècniques i algoritmes varis que serveixen per identificar automàticament símbols o caràcters, que pertanyen a un determinat alfabet, a partir de la imatge en qüestió per emmagatzemar-ho en forma de dades.

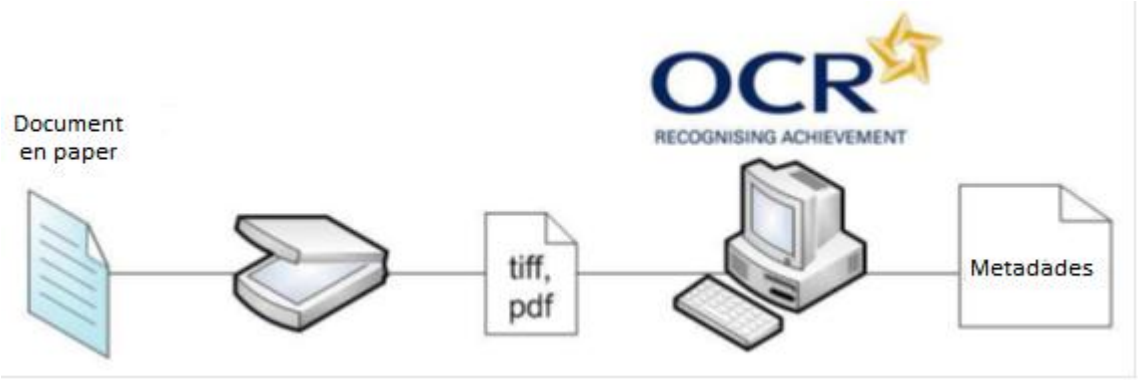


Figura 1. Reconeixement de caràcters de documents en paper mitjançant OCR

Aquest procediment consta de quatre etapes:

3.2.1 Binarització

La major part d'algorismes d'OCR parteixen de la base d'una imatge binaria (dos colors) per tant és convenient convertir una imatge amb escala de grisos, o color, en una en blanc i negre, de tal manera que es preservin les propietats essencials de la imatge. Una forma de fer-ho es mitjançant l'histograma de la imatge on es mostra el número de píxels per cada nivell de gris que apareix a la imatge. Per binaritzar-la s'ha de triar un llindar adequat, a partir del qual tots els píxels que no el superin es convertiran en negre i la resta en blanc.

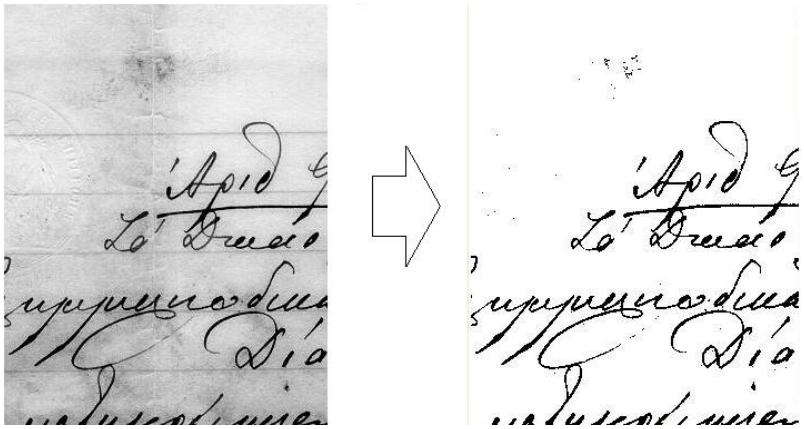


Figura 2. Aplicació de la binarització d'una imatge

Mitjançant aquest procés obtenim una imatge en blanc i negre on queden clarament marcats els contorns dels caràcters i símbols que conté la imatge.

3.2.2 Segmentació de la imatge

Consisteix en detectar i aïllar les regions de la imatge que contenen text. Aquesta segmentació es pot dur a terme a partir de diversos procediments, ja que no hi ha un mètode genèric que sigui suficientment eficaç per l'anàlisi d'un text. Tot i això, les tècniques més utilitzades són variacions dels mètodes basats en projeccions lineals. El procés aconsegueix descompondre la imatge en diferents entitats lògiques suficientment significatives pel seu reconeixement.

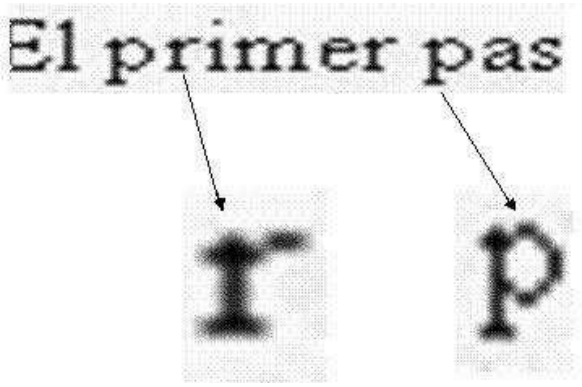


Figura 3. Segmentació de la imatge

3.2.3 Simplificació dels components

Aquesta tècnica té la finalitat de simplificar la forma de les dades que es desitgen capturar, facilitant així, el seu reconeixement. Per fer-ho, s'eliminen successivament els punts dels contorns de cada component, mantenint les proporcions originals i la seva tipologia.

Figura 4. Simplificació dels components

3.2.4 Comparació amb patrons

Finalment, es comparen les formes obtingudes anteriorment amb patrons emmagatzemats en bases de dades per obtenir el caràcter a representar.

Figura 5. Comparació de la imatge amb patrons coneguts

3.3 La gestió documental

La gestió documental és la responsable d'administrar el flux de documents en una empresa, a més de custodiar de forma segura i facilitar la consulta dels mateixos de forma ràpida i eficaç. Addicionalment, pot permetre l'ús de *workflows* o fluxos de treball que automatitzen algunes tasques com la validació de factures o altres documents, fent el procés molt més àgil. D'aquesta manera, queden definides dues funcionalitats principals:

3.3.1 Custòdia segura

Els documents han de quedar emmagatzemats de forma segura en una base de dades. Aquesta, contemplarà funcionalitats com la de definir el temps que es mantindrà un document o els sistemes de còpies de seguretat. A més, permetrà la consulta dels documents emmagatzemats només a usuaris degudament autoritzats.

3.3.2 Workflows i col·laboració

Un gestor documental permet definir el flux de transmissió dels documents, que pot comprendre la creació d'una factura, la seva validació per part d'un usuari i l'aprovació o revisió d'un supervisor que, en cas de no aprovar, podrà tornar la factura a l'usuari validador de forma automàtica perquè corregeixi qualsevol possible error. També hi ha la possibilitat que un document sigui realitzat per més d'una persona. Donat aquest cas, l'aplicació proporciona eines perquè les actualitzacions que es fan sobre un document no s'alterin mútuament. A més, els gestors documentals solen incorporar un control de versions, que permet tornar a estats anteriors dels documents de forma

senzilla. És possible crear una gran varietat de fluxos de treball perquè s'adeqüin als diferents tipus de documents amb els que el gestor documental ha de treballar.

4. Per què és necessari un bon procés documental?

Un sistema automatitzat de qualsevol procés de negoci és molt probable que produeixi dades més netes, resultant en informació de major qualitat, menor maneig d'excepcions, i una millor gestió de processos.

Dins la funció empresarial les factures segueixen sent un document pesat. Gairebé el 77% de totes elles segueixen sent en paper. Aquesta és una àrea que encara està madurant i hi ha una amplia àrea de recerca en termes d'automatització de processos i adopció de tecnologia, amb gran potencial per impulsar l'estalvi de costos en l'administració de les factures entrants d'una organització.

El problemes principals que es presenten en grans organitzacions són el cost, el temps i la manca de visibilitat en l'administració d'aquest tipus de documents. S'ha demostrat que aquests aspectes es poden millorar substancialment amb l'automatització del procés mitjançant la tecnologia digital actual.

Basant-nos en l'estudi *Invoicing and Workflow*[ii] realitzat per *Aberdeen Group*[iii], una consultora especialitzada en la recerca de processos empresarials i tecnològics, podem fer-nos una idea del cost que té el procés d'administració d'una factura per una companyia i la possible reducció de costos si s'inverteix en l'automatització d'aquest procés.

	Cost i temps en el procés d'una factura	Factures elegibles a descompte	Cost anual per 120.000 factures
Les més eficients	3,34 $ 4,1 dies	90%	400.800 $
La mitjana d'empreses	6,29 $ 6,1 dies	47%	754.800 $
Les menys eficients	6,67 $ 16,3 dies	18%	2.000.400 $

Taula 1. Cost d'administració de les factures. *Aberdeen Group,* Abril 2011

En el següent gràfic del mateix estudi podem veure les estratègies més importants que havien seguit les empreses auditades per millorar el procés i podem comprovar que les que figuren dins del grup de les més eficients havien invertit d'una manera significativament superior a les demés en l'automatització de la captura de factures i processos de flux de treball relacionats.

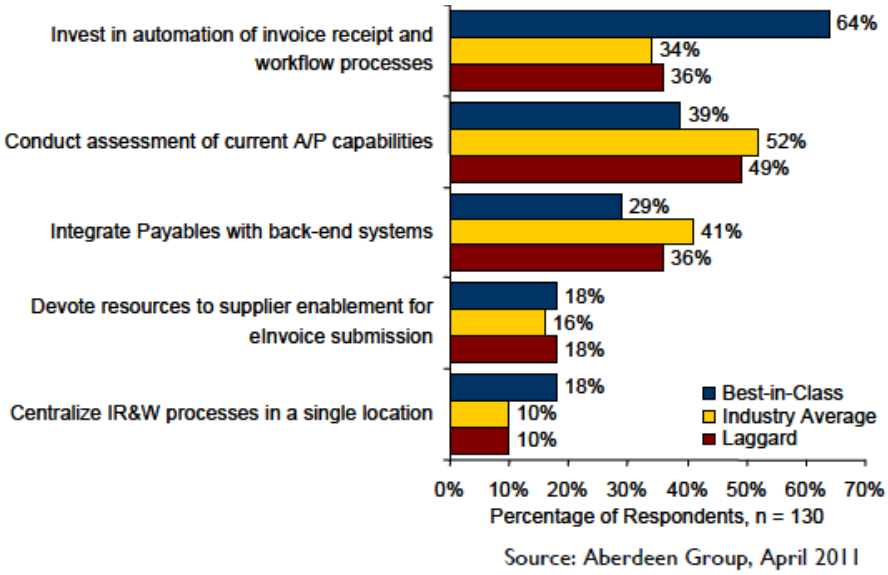


Figura 6. Gràfic de les estratègies més importants seguides per millorar l'administració de factures.

***Aberdeen Group,* Abril 2011**

A més, si a la millora d'aquests processos afegim el de la digitalització certificada també podem aconseguir optimitzar altres aspectes com són els costos de l'emissió i la recepció del document en paper:

EMISSIÓ			
Paper	**cost / unitat**	**Factura electrònica**	**cost / unitat**
Impressió	0,12	Hardware & Software	0,01
Enviament (sobre, segell)	0,29	Tràfic	0,03
Tractament manual	0,35	Gestió (Dpt. Administració)	0,02
Total	**0,77**	**Total**	**0,06**
Estalvi per factura = 0,70 €			

Taula 2. Benefici de l'emissió de factures electròniques. *Agencia Española de Codificación Comercial* (AECOC)

RECEPCIÓ			
Paper	**cost / unitat**	**Factura electrònica**	**cost / unitat**
Recepció i manipulació	0,07	Hardware & Software	0,01
Gravació	0,15	Tràfic	0,03
Tractament manual	1,68	Gestió (Dpt. Administració)	0,02
Arxivament	0,97	Arxivament	0,03
Total	**2,87**	**Total**	**0,09**
Estalvi per factura = 2,78 €			

Taula 3. Benefici de la recepció de factures electròniques. *Agencia Española de Codificación Comercial* (AECOC)

Les solucions de processament de factures i automatització de comptes a pagar optimitzen el procés de captura de factures per augmentar la velocitat de processament i reduir els costos generats per l'entrada de dades. A més, permeten aprofitar els *workflows* de millors pràctiques per llançar els processos de revisió, codificació i aprovació de factures, el que permet un ràpid retorn de la inversió.

Aquests són alguns dels beneficis més importants que es poden assolir:

- ✓ **Augment de la velocitat de processament i reducció dels costos de processament**

Amb la captura automàtica de factures electròniques i en paper tan bon punt es reben s'aconsegueix l'eliminació dels retards generats per l'entrada de dades gràcies a l'ús del reconeixement automàtic. Això permet el llançament automàtic de fluxos de treball d'aprovació de factures per a la seva codificació i aprovació adequades, el que permet realitzar els pagaments a temps. Reducció dels costos d'entrada de dades gràcies al reconeixement i la validació automàtics de les línies de detall de les factures. Es calcula que la recuperació de la inversió es materialitza en menys de 12 mesos.

- ✓ **Optimització de la visibilitat i millora del control**

Amb la conversió de totes les factures en arxius i dades electròniques en format estàndard, qualsevol persona autoritzada, o fins i tot diverses al mateix temps, pot accedir des de l'escriptori de l'ordinador a aquestes dades en qualsevol moment. També es pot realitzar la implementació d'un procés formal i auditable per capturar factures, independentment del format o contingut, i de mesures de seguretat i polítiques d'emmagatzematge per reduir el risc d'incompliment de la normativa.

- ✓ **Millora de la qualitat de les dades i de la precisió**

Com que es validen totes les factures a mesura que són capturades els processos d'aprovació s'executen sense demora i es poden millorar les decisions de negoci gràcies al subministrament de dades de proveïdors d'alta qualitat.

- ✓ **Gestió optimitzada de tresoreria i gestió del rendiment de comptes a pagar en temps real**

Es poden aprofitar al màxim els descomptes per pagament i evitar les penalitzacions per retards. A més, s'obté visibilitat sobre el procés de comptes a pagar per garantir que es respecten els indicadors clau de rendiment (*KPI*, en anglès) i els acords de nivell de servei (*SLA*, en anglès).

5. Visió general i presentació de la solució

A continuació es descriu l'abast, l'arquitectura, les plataformes software i el circuit procedimental que vol assolir aquest projecte.

El procés comença amb l'entrada dels documents a digitalitzar i acaba amb les dades d'aquests documents a l'ERP de l'empresa. La solució que es proposa contempla l'extracció intel·ligent de les dades i les metadades requerides pel client, la validació de les mateixes i la transferència dels arxius i dades necessaris per la comptabilització de les factures a un sistema ERP.

S'ha de destacar que el software ERP a utilitzar no queda lligat a la solució, cosa que facilita la implantació d'aquesta a diferents entorns. A més, permet a les empreses seguir utilitzant el sistema que estiguin utilitzant, ja que sovint aquests abasten una gran complexitat i una gran inversió que fan que sigui difícil de modificar o adaptar a altres sistemes.

A l'hora d'escollir la plataforma per realitzar el projecte, només s'han pres en consideració aquelles amb les que treballa l'empresa TBS en l'àmbit del qual s'ha realitzat aquest Projecte de Fi de Carrera. A més, per raons tècniques, l'elecció de les tecnologies amb les que s'ha treballat ha respòs a factors estratègics i comercials propis de TBS.

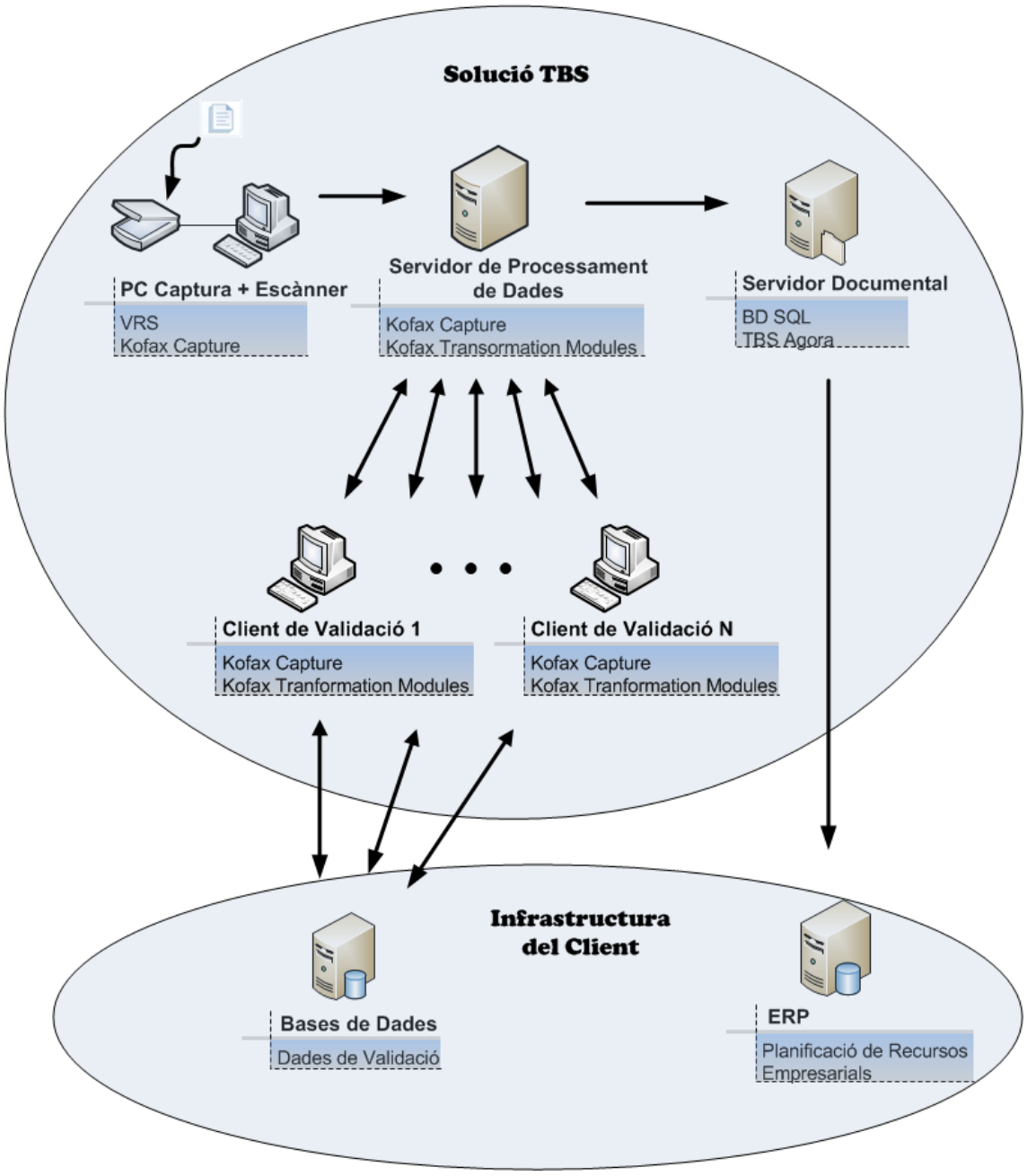


Figura 7. Diagrama general de l'arquitectura

5.1 Plataforma de Captura Digital

TBS és partner d'un dels fabricants de software més importants del món en l'àmbit de la captura digital de documents. Aquest és Kofax, que proporciona un software sofisticat de captura i extracció digital, el qual és programable i el podem adaptar a les nostres necessitats. A més, Kofax va ocupar el primer lloc de las 3 categories principals del "The Forrester Wave™: Multichannel Capture, Q3 2012"[iv], en aquesta matèria:

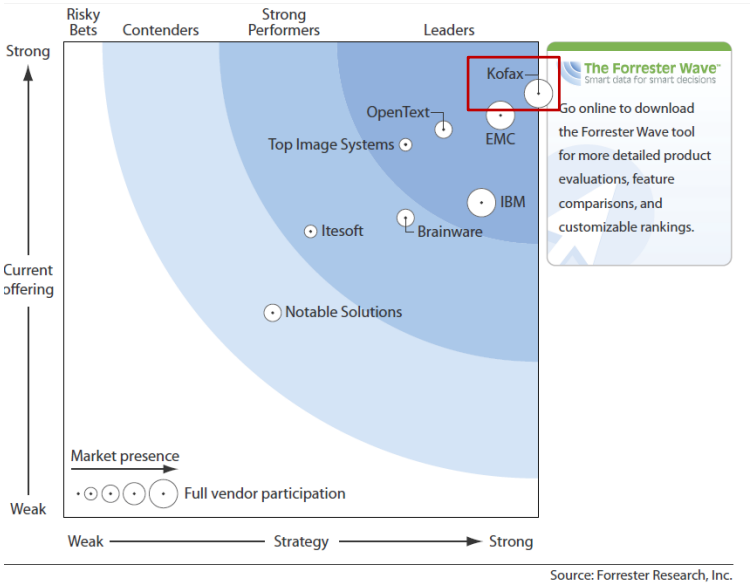


Figura 8. The Forrester Wave™: Multichannel Capture, Q3 2012. *Forrester Research, Inc.*

És per això que s'han descartat altres possibilitats, ja que no oferien ni els mateixos beneficis comercials ni les mateixes oportunitats tan a nivell de negoci com a nivell tecnològic.

La solució que contempla aquest projecte comprèn els productes Kofax VRS, Kofax Capture i Kofax Transformation Modules, que impliquen els processos de Capturar, Transformar i Entregar documents.

Figura 9. Esquema de funcionament de Kofax

5.1.1 Kofax VRS (Versió Elite)

Kofax VRS (Virtual ReScan) és una aplicació que millora substancialment el procés de digitalització per obtenir de les imatges escanejades una qualitat òptima i poder extreure les dades contingudes en elles de manera més eficaç. Aquesta aplicació és configurable i d'entre les possibilitats que té s'inclouen les següents:

- Orientació automàtica de documents: Detecta si el document escanejat està inclinat o del revés i corregeix la situació per tal de facilitar l'extracció posterior.
- Detecció i eliminació de pàgines en blanc: Així ens estalviem pàgines a processar innecessàries.
- Retall automàtic de documents: Detecta si el document és d'una mida diferent o si hi ha parts que no pertanyen al document digitalitzat i les elimina.
- Eliminació de marques: Pot ser que al digitalitzar un document li hagin quedat marques com forats o grapes, el software les elimina per tal de que no contribueixin al soroll de l'extracció.
- Neteja intel·ligent: Intenta corregir qualsevol efecte que dificulti la nitidesa de la imatge obtinguda com podria ser documents arrugats o brutícia de fotocòpies.
- Gestió de perfils: Configuració de perfils en funció del document a escanejar i de l'escàner.

Així, el software ens permet preprocessar o reescanejar, com el seu nom indica, la imatge per tal de facilitar la tasca posterior de reconeixement amb OCR i evitar tots els elements que en puguin crear confusió a l'hora de tractar amb l'extracció intel·ligent de dades de la imatge. A més, val a dir que no té cap impacte en la velocitat de digitalització marcada per l'escàner i és compatible amb la gran majoria de dispositius professionals que hi ha al mercat.

5.1.2 Kofax Capture (Versió 10.0 SP1)

Kofax Capture és l'eix principal a partir del qual s'afegeixen els diferents mòduls de Kofax. Aquesta aplicació permet la captura, indexació i validació de documents. Tot i això, per aquest projecte s'utilitza de forma conjunta amb Kofax Transromation Modules (KTM), ja que aquest últim ens amplia substancialment la capacitat de captura i validació de Kofax Capture, com es veurà a continuació.

Aquesta aplicació també ens permet muntar el sistema de captura i validació de documents en una estructura de client-servidor, de manera que podem escanejar els documents des de múltiples estacions, però extretes i processades en un únic punt, podent validar les dades des de l'estació on s'ha realitzat la digitalització o des d'altres equips que només faran de clients de validació.

5.1.3 Kofax Transformation Modules (Versió 5.5 SP2)

Les funcions de l'aplicació són la classificació de tot tipus de documents, l'extracció de les seves dades i la validació d'aquestes.

És aquest mòdul de Kofax el que ens permet optimitzar aquests processos partint de regles lògiques i programables. Això ho fa a través de plantilles personificables per cada tipus de document i, a més, és capaç "d'aprendre" guardant en memòria els resultats processats de mostres anteriors que presentin la mateixa estructura. KTM ofereix un alt grau de personalització i permet crear formularis de validació de documents amb tot tipus de funcionalitats. Per fer-ho, fa ús d'un entorn de desenvolupament en WinWrap Basic[v], un llenguatge basat en el llenguatge Visual Basic. Gràcies a aquest entorn podem programar i integrar totes les funcionalitat amb les que podrem validar les factures.

5.2 El gestor documental

El gestor documental TBS Agora és un entorn web que ens permet explotar tota la informació extreta de la digitalització a través d'una interfície amigable per l'usuari. A partir de les dades emmagatzemades en una base de dades TBS Agora permet consultar-les, modificar-les i indexar-les de manera que l'usuari pugui fer-ne un ús eficient i optimitzar els processos de negoci als quals estan associades. El gestor també ens dóna la possibilitat de crear fluxos de treball pels quals la documentació digitalitzada podrà ser visualitzada o aprovada en funció del tipus d'usuari que aparegui com actor del sistema.

La creació del gestor documental està desenvolupat per l'empresa TBS i no forma part d'aquest PFC, tot hi això s'ha tingut en compte com a part del circuit que té el document des de la seva digitalització fins la seva entrada a l'ERP.

5.3 Circuit de la solució

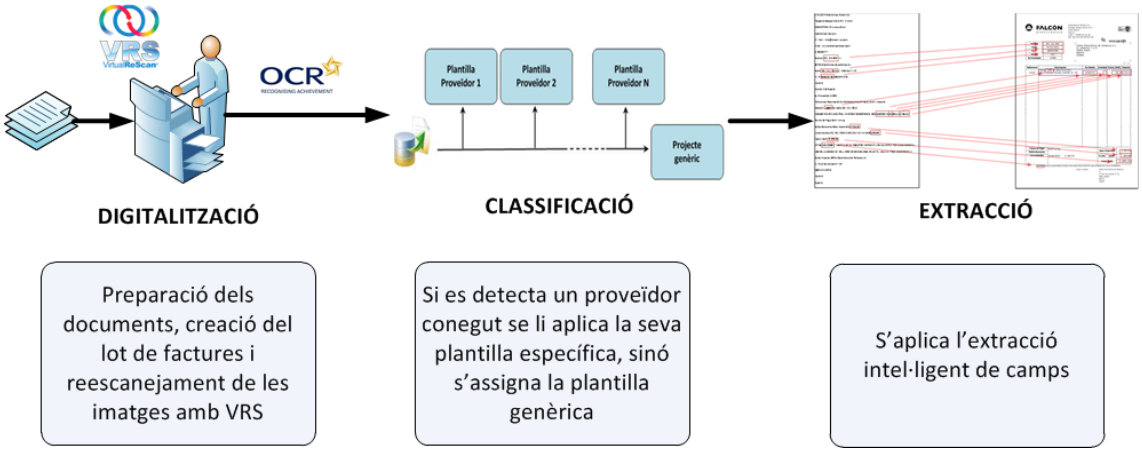
El circuit del sistema comença amb l'usuari preparant els documents que s'han de digitalitzar. Per fer-ho, en el cas que es tracti de documents de varies pàgines l'usuari encarregat de l'equip de captura ha de col·locar un separador entre cada document per tal d'indicar al sistema on acaba un document i comença el següent. Aquests documents en paper es digitalitzen a partir d'un escàner connectat a l'estació de captura on hi ha instal·lat el software de captura de Kofax. Primerament, l'usuari ha de crear un conjunt de factures, que a partir d'ara anomenarem lot, que contindrà els documents que es disposa a digitalitzar. Al crear aquest lot l'usuari ha de definir dues propietats: escollir la societat a la que van dirigides les factures i si aquestes són factures amb comanda o sense. Aquestes són propietats que influiran en el tractament del lot com es veurà més endavant.

Un cop creat el lot l'usuari ha de polsar el botó de digitalitzar i a més de capturar-se digitalment tots els documents introduïts a l'escàner, s'executa el mòdul VRS que aplica el preprocessament a les imatges obtingudes per tal d'optimitzar l'extracció. A continuació, aquestes imatges són enviades al servidor de processament on s'aplica l'algorisme OCR configurat. En aquest punt el servidor executa un circuit lògic per tal d'aconseguir les dades més fiables possibles i validar-les en la mesura del possible, i en el cas de que no pugui ser així, mostrar que les dades que l'usuari visualitza no són fiables. Aquest circuit lògic consta de quatre etapes diferenciades:

- Classificació: El document s'assigna a una plantilla definida. En el nostre cas, definim una plantilla genèrica on es realitzarà l'extracció intel·ligent i el tractament general de les dades. Dins d'aquesta plantilla la factura es podrà assignar a una subplantilla, que podrà ser la pròpia d'un proveïdor.
- Extracció: Per obtenir les dades del document i assignar-les als camps corresponents utilitzem localitzadors de patrons de caràcters tan espacials com

lògics, així com regles de negoci, tenint en compte el resultat de l'OCR i la imatge obtinguda.

- Formateig: Per tots els camps que s'obtinguin de l'extracció i que tinguin un format conegut o no aleatori es programa una solució per complir amb el patró que pugui seguir.
- Validació automàtica: Amb els camps formatats es procedeix a fer totes les comprovacions possibles per determinar si els resultats obtinguts són els reals i, en cas contrari, corregir-los amb les dades que sabem que si que són vàlides. Aquesta és la que s'ha anomenat validació automàtica per diferenciar-la de la validació de camps que farà l'usuari. Aquesta validació es portarà a terme a partir de bases de dades i regles de negoci conegudes aplicables a les factures.



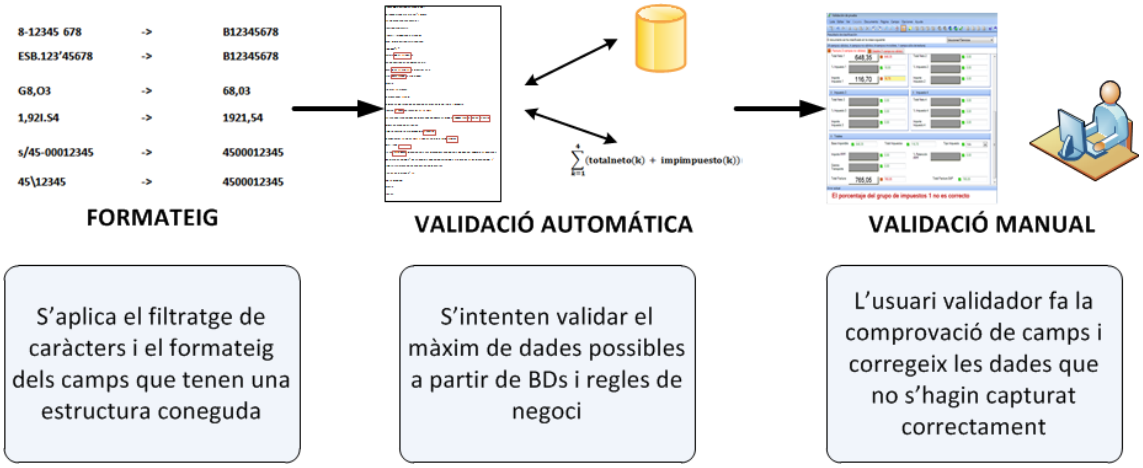


Figura 10. Esquema explicatiu del circuit de la solució

Un cop processada la imatge es mostra a l'usuari un formulari de validació al costat de la imatge digitalitzada on aquest podrà comprovar si tots els camps han sigut extrets amb èxit o introduir-los en cas que no sigui així. Al validar el document, les dades són enviades al servidor de base de dades del qual el gestor documental TBS Agora obtindrà les dades per mostrar-les al client mitjançant l'entorn web. Finalment, serà el gestor documental l'encarregat d'introduir les dades al ERP perquè es comptabilitzi la factura al sistema d'informació de la corporació.

5.4 Diagrama de flux

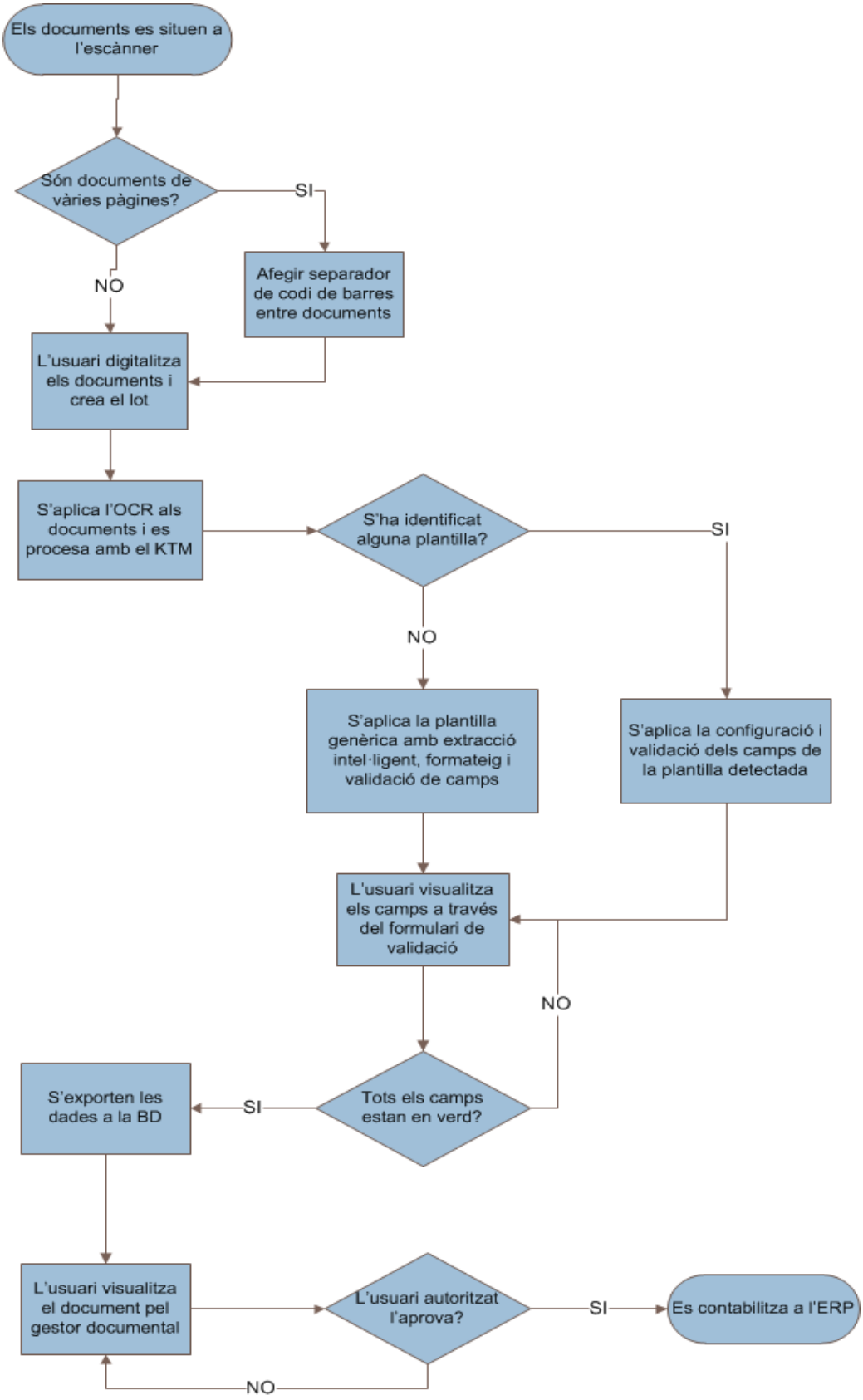


Figura 11. Diagrama de flux

6. Anàlisi de requisits

L'objectiu del sistema a desenvolupar i implantar consta de varies etapes de processament de factures i un dels objectius que es pretén assolir dins del circuit és que l'usuari hagi d'interaccionar el mínim possible amb l'aplicació, però amb la màxima seguretat de que les dades finalment emmagatzemades siguin dades vàlides, estiguin al document original o no. D'aquesta manera aconseguirem l'automatització i optimització desitjats del procés.

6.1 Característiques dels usuaris

Dins del sistema de captura es requereixen dues interaccions directes per part del usuaris que seran realitzades per dos perfils diferents:

- L'usuari del punt de captura: Aquest usuari serà l'encarregat de col·locar els documents de manera adequada a l'escàner perquè aquest pugui detectar-los correctament. També és el que realitzarà l'acció de digitalitzar aquests documents separant-los en lots de factures.
- L'usuari validador: Aquest usuari serà l'encarregat de revisar i validar els camps dels documents un cop aquests hagin estat processats pel sistema. Per la funció que ha de desenvolupar el perfil d'aquest usuari ha de pertànyer a l'àrea administrativa i/o econòmica de la corporació, ja que ha de tenir coneixements sobre documents de caire financer com són les factures.

6.2 Característiques de la plataforma final

Aquest projecte va dirigit a empreses amb un gran volum de facturació, pel que es sobreentén que ja tenen una infraestructura tecnològica implantada. Per això, aquest sistema s'adapta per integrar-se amb els sistemes operatius més habituals en aquest tipus d'empreses pel perfil d'usuaris que hauran d'interaccionar amb l'aplicació, és a dir, *Microsoft Windows*. A més, l'empresa proveïdora del software principal del que consta el projecte, Kofax, és *partner* de l'empresa Microsoft, pel que podrà donar

suport fàcilment a incidències que puguin sorgir derivades de la interacció de l'aplicació amb el sistema operatiu.

6.3 Funcionalitats

Les funcionalitats essencials del sistema són:

- Captura intel·ligent de les dades de les factures introduïdes al sistema requerides per la Corporació Albatros.
- Formateig i validació automàtica d'aquestes dades.
- Possibilitat de l'usuari de modificar les dades mitjançant un formulari i la imatge de la factura.
- Possibilitat de l'usuari de cercar dades contingudes en bases de dades.
- Exportació automàtica de les dades finals a una base de dades relacional.
- Certificació digital automàtica de les factures.
- Implantació d'un gestor documental on es podran dur a terme diferents fluxos de treball amb els documents.
- Vinculació automàtica amb l'ERP corporatiu.

7. Desenvolupament de la solució de captura

7.1 Camps a capturar

En primer lloc, i per poder determinar la lògica estructural del tipus de document a tractar, s'ha de disposar d'un volum de mostres que sigui suficient per poder desenvolupar el projecte en qüestió. En aquest cas, la corporació Albatros ha facilitat **266** factures de **109** proveïdors diferents.

El següent pas és decidir en funció de les mostres i de la voluntat del client quines són les dades i les metadades que volem extreure de les factures i determinar així, els diferents camps que hem de crear tan al projecte de Kofax com a la base de dades on es desaran finalment. Els camps que volem obtenir de cada una de les factures són:

Nom del camp	Descripció
Numfac	Número de la factura
Fechafac	Data de la factura
Numped	Número de comanda
Numalb	Número d'albarà
Codprov	Codi SAP del proveïdor que servirà per identificar-lo posteriorment al ERP
Cifprov	CIF del proveïdor
Nomprov	Nom del proveïdor
Paisprov	País del proveïdor
Tipofacsap	Tipus de factura SAP: MM o FI
Tipofac	Identifica el document com a factura o abono
totalneto0	Import de la factura exempt d'impostos
pctimpuesto1	Percentatge de l'impost 1
impimpuesto1	Import de l'impost 1
totalneto1	Import total on s'ha aplicat l'impost 1
pctimpuesto2	Percentatge de l'impost 2
impimpuesto2	Import de l'impost 2
totalneto2	Import total on s'ha aplicat l'impost 2
pctimpuesto3	Percentatge de l'impost 3
impimpuesto3	Import de l'impost 3
totalneto3	Import total on s'ha aplicat l'impost 3
pctimpuesto4	Percentatge de l'impost 4
impimpuesto4	Import de l'impost 4
totalneto4	Import total on s'ha aplicat l'impost 4
Pctirpf	Percentatge de IRPF
Impirpf	Import de IRPF a aplicar
Baseimp	Base imposable total de la factura
totalpimpuestos	Suma total dels impostos de la factura
totalfactura	Import total de la factura
Codtrans	Codi de transferència del document que ens

	servirà per determinar l'estat de la factura a la base de dades. Per defecte "00" (pendent de transferir)
Moneda	Moneda de la factura
Detalle	Taula de línies de comanda de la factura

Taula 4. Llista de camps a capturar

Camps de la taula de línies de comanda anomenada "Detalle":

dt_posicion	Posició de comanda per línia de factura
dt_numped	Número de comanda per línia de factura
dt_numalb	Número de albarà per línia de factura
dt_articulo	Codi d'article per línia de factura
dt_descripcion	Descripció de l'article per línia de factura
dt_cantidad	Quantitat de l'article per línia de factura
dt_preciounitario	Preu unitari de l'article per línia de factura
dt_preciototal	Preu total per línia de factura

Taula 5. Llista de camps de la taula "Detalla" corresponents a les línies de comanda

Amb aquests camps definits ja podem crear el projecte al "Project Builder", el mòdul de KTM on s'han de definir els camps i desenvolupar el codi per la classificació, l'extracció, el formateig i la validació de cada un d'ells, així com la construcció del formulari de validació que es mostrarà a l'usuari. També s'han creat els mateixos camps al projecte de Kofax Capture per tal de poder fer la sincronització i exportació de les dades dels camps provinents de KTM cap a la base de dades.

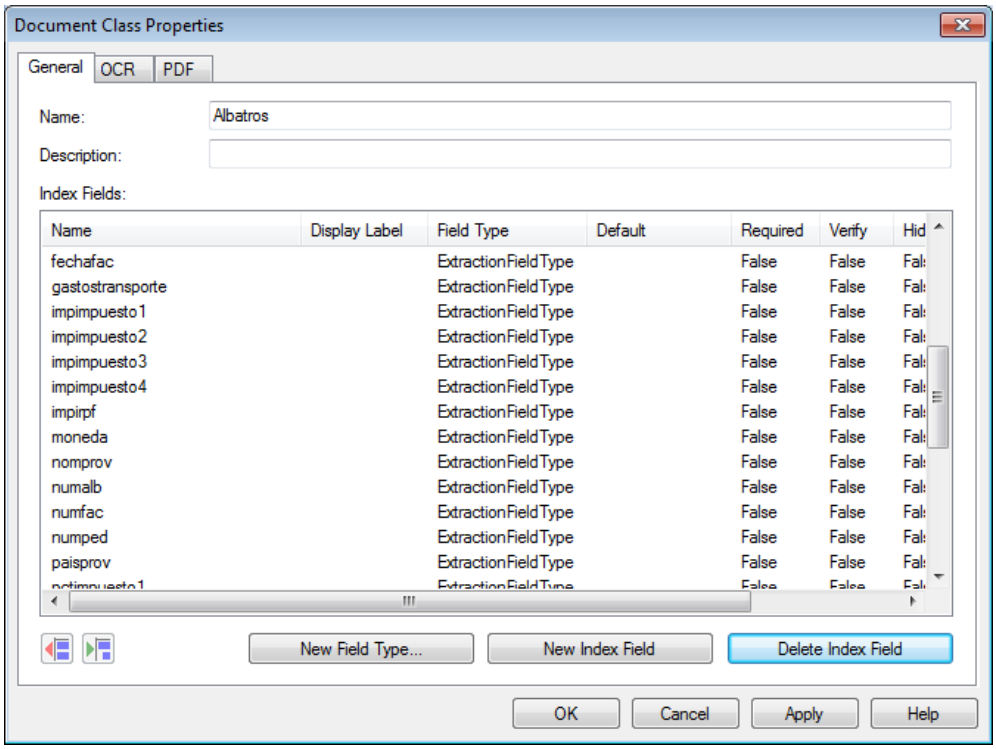


Figura 12. Finestra de Kofax Capture on es mostren la llista de camps creats

També definim els paràmetres d'entrada que haurà de seleccionar l'usuari al crear el lot de factures a escanejar, és a dir, la classe de lot (d'una o vàries pàgines), la societat de la factura i el tipus, si és una factura amb comanda (MM) o una factura sense comanda (FI).

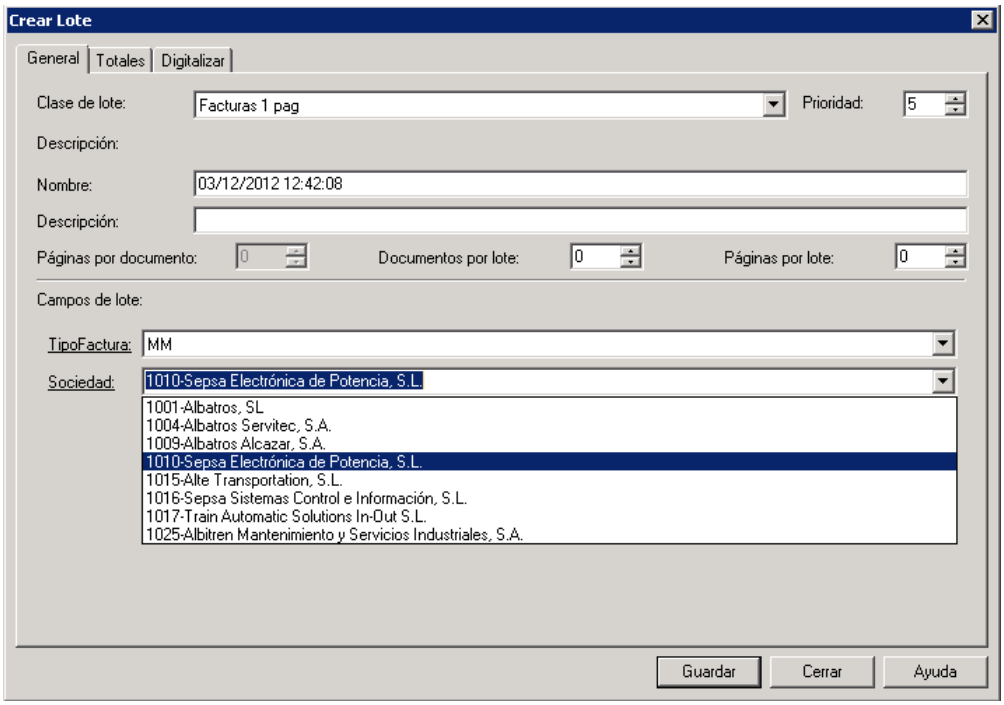


Figura 13. Creació d'un lot a través de Kofax Capture

7.2 Separadors

7.2.1 Separadors de documents

Una problemàtica que se'ns presenta és com determinar el número de pàgines pels documents escanejats que tenen més d'una pàgina, ja que l'objectiu de la solució és que l'usuari pugui escanejar tots els fulls dels documents sense haver d'interaccionar amb el dispositiu d'escaneig més que a l'hora d'executar l'acció de digitalització. Aquest problema s'ha resolt amb intèrprets de codis de barres. Abans d'inserir aquest tipus de documents a l'escàner col·loquem un full amb un codi de barres que conté un nombre i format coneguts a l'inici de cada document. D'aquesta manera cada cop que l'aplicació detecti aquest codi de barres interpretarà que s'està processant un altre document.

Figura 14. Separador de documents

7.2.2 Separadors de documentació adjunta

Sovint les factures arriben amb d'altra documentació, com poden ser albarans, copies de la mateixa factura o simplement les condicions financeres del proveïdor. És documentació que l'empresa vol conservar amb la factura, ja que pot ser útil en algun moment, però que no només no té incidència en el projecte que es duu a terme per captar les dades de les factures, sinó que pot dificultar el reconeixement digital d'aquestes dades afegint a vegades varies pàgines que poden confondre el programa a l'hora d'executar l'OCR i l'extracció dels camps. A més, s'ha de tenir en compte que la llicència del software de Kofax permet un número limitat de documents l'any, pel que a l'empresa client li és molt interessant poder descartar tota aquesta documentació adjunta que no necessita OCR perquè no es processi amb les factures normals i no compti dins d'aquets número de documents contractats.

Per poder mantenir aquest tipus de documentació adjunta a la factura s'ha definit un altre separador amb un altre codi de barres conegut diferent del separador de documents. El reconeixement d'aquest codi de barres s'ha desenvolupat perquè s'executi just abans d'executar l'OCR. Per tant, quan el programa detecti aquest

separador descartarà totes les pàgines que vinguin a continuació ometent el processament OCR i, per tant, tota la resta del procés pel que passa el document.

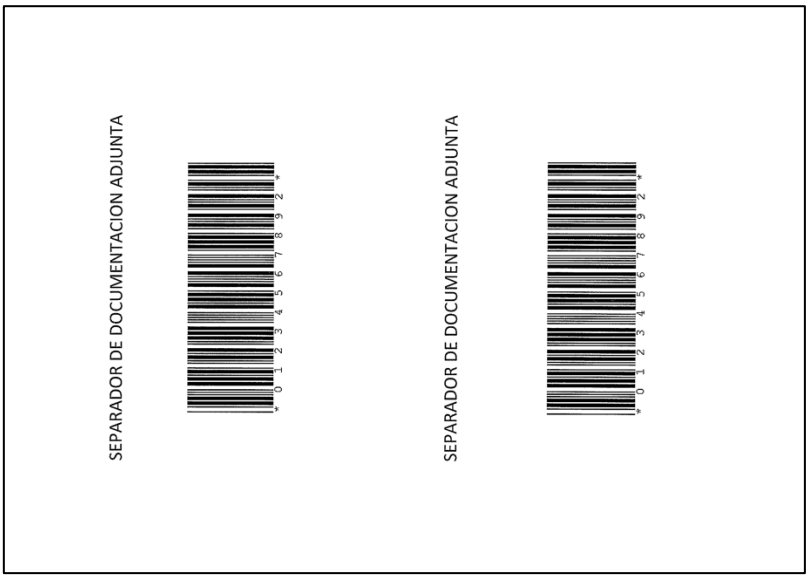


Figura 15. Separador de documentació adjunta

7.3 OCR

El procés d'OCR s'encarrega de transformar el document obtingut en forma de fitxer d'imatge en un fitxer de text. Aquest procés depèn en gran mesura de la qualitat de la imatge obtinguda, i és per això que la correcta configuració de VRS té una gran importància. S'ha decidit usar el motor OCR *Finereader*[vi] ja que, després de proves exhaustives és el que millors resultats ha donat respecte a l'altra alternativa que es tenia en compte, que era *Recostar*[vii].

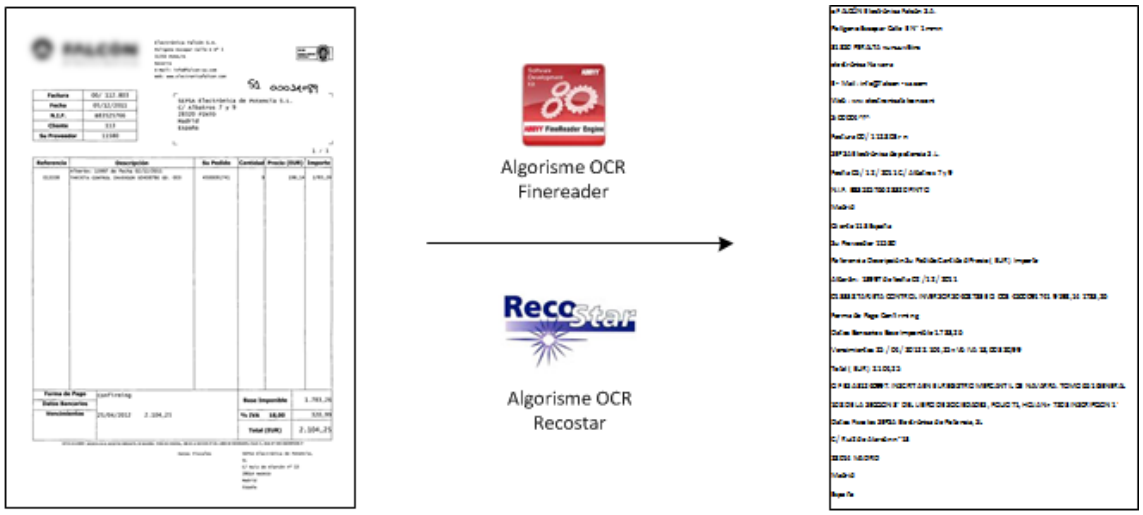


Figura 16. Aplicació de l'OCR

Després de realitzar proves amb un número important de factures s'ha determinat que hi ha tipus de documents en els quals no cal aplicar el reconeixement de caràcters a totes les pàgines, ja que a més d'afegir informació innecessària al document de text que pot influir negativament a l'hora de fer l'extracció intel·ligent dels camps, es consumeixen pàgines processades de la llicència, tal i com passa amb la documentació adjunta. Aquests són els documents de vàries pàgines sense línies de comanda. Per aquests documents s'ha inhabilitat l'OCR de totes les pàgines excepte la primera i la última, tenint en compte que és en aquestes pàgines on figuren les dades dels camps que volem capturar.

7.4 Tractament de les dades

A continuació, s'explica el funcionament de les diferents fases per les que passen els documents. Per cada fase es descriuen com s'han desenvolupat alguns dels camps a capturar que són representatius dels tres grups en que s'han agrupat: el CIF del proveïdor, dels camps de la capçalera de la factura; el total de la factura, del grup de camps dels imports; el número de comanda i les línies de detall d'aquesta.

7.4.1 Classificació

Com que el que volem és definir l'extracció de dades per un tipus de document semiestructurat, com és el cas de les factures, el que farem primer de tot és crear una plantilla genèrica que contingui les configuracions i el codi que s'executarà per defecte en el cas que no s'hagin detectat altres plantilles. Aquesta plantilla funcionarà a mode de classe, de manera que podrem crear subplantilles (subclasses de documents) que esdevindran tipus de factures que podrem definir i desenvolupar de manera concreta per un proveïdor en particular. Tot i això, en una empresa gran com amb la que treballem, la solució que li és més interessant és aquella en que és processin amb el major encert possible la major quantitat de factures sigui quina sigui la seva estructura, ja que és inviable crear una plantilla per cada proveïdor. Partint d'aquesta premissa s'ha desenvolupat la plantilla genèrica amb la intenció d'abastar el màxim d'estructures de factures possibles i posteriorment, s'ha sotmès el projecte a un període de proves amb les mostres subministrades per la corporació client. En funció dels resultats d'aquestes proves i del volum de cada tipus de factura dins la corporació (hi ha proveïdors més importants que d'altres) s'ha pogut valorar quines subplantilles es poden afegir i de quina manera per cada proveïdor.

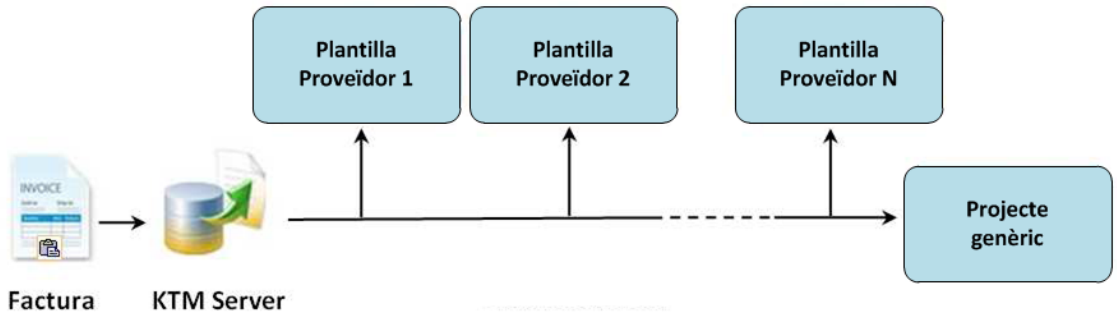


Figura 17. Procés de classificació a través de KTM

Hi ha dues maneres possibles de classificar un document:

- Classificació per disseny: Es digitalitzen varies mostres de cada proveïdor i s'assignen a la plantilla que li correspondrà. A partir d'aquestes imatges KTM crea l'estructura de la plantilla d'acord amb l'estructura de les factures i cada cop que es processi un document amb aquella estructura el classificarà amb la plantilla assignada.
- Classificació per contingut: S'han de triar una o més cadenes de caràcters perquè identifiquin cada plantilla. Aquestes cadenes de caràcters anomenades "frases" es busquen al fitxer de text resultant de l'execució de l'OCR podent definir un percentatge de confiança per cada frase. Perquè funcioni correctament idealment aquestes frases haurien de ser úniques.

El mètode que s'ha triat per classificar un document en una subplantilla ha sigut el de classificació per contingut, ja que al ser un document semiestructurat, però sovint amb estructures molt semblants, el programa pot confondre la identificació de la plantilla amb facilitat si es classifica la factura mitjançant la imatge. Després d'unes quantes proves amb els dos mètodes s'ha decidit que la classificació per contingut és la més fiable.

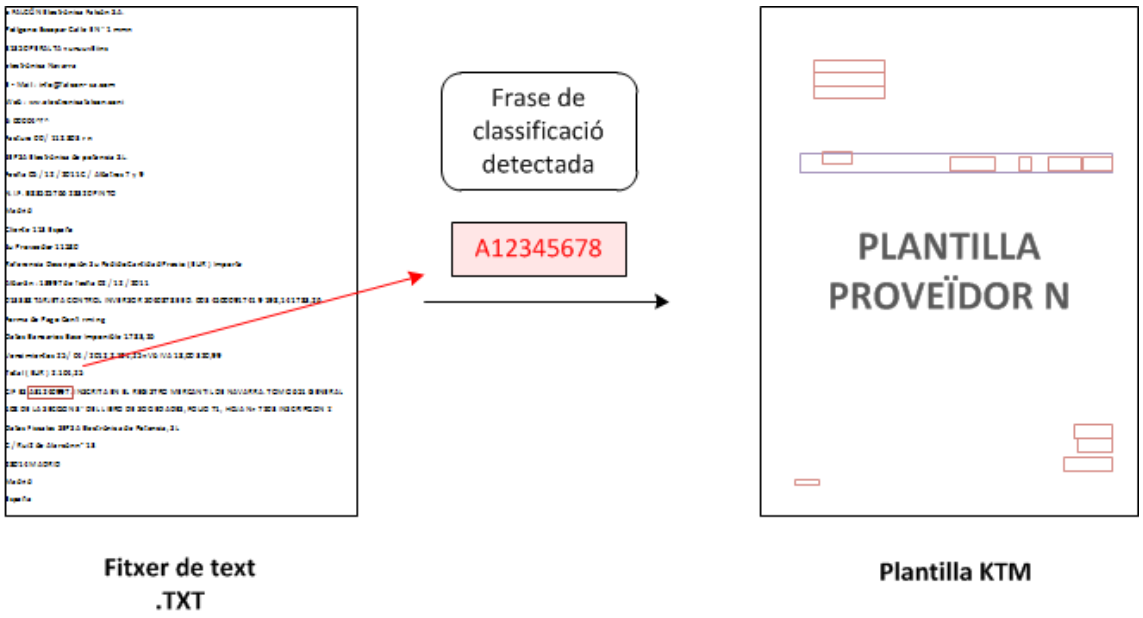


Figura 18. Classificació d'una plantilla a KTM

Tenint en compte les característiques de la classificació per contingut la frase ideal per identificar un proveïdor és el seu CIF, ja que aquest és únic. Tot i això, s'han afegit altres opcions com la pàgina web o el número de telèfon, ja que s'ha detectat que en moltes ocasions el CIF figura en una grandària molt reduïda i la captura no sempre és 100% correcte.

7.4.2 Extracció

El software de KTM ens ofereix uns localitzadors, la funció dels quals és la de localitzar un cert camp al document. Aquests localitzadors es basen en la configuració del patró lògic que segueix cada cadena de caràcters que volem detectar, en la seva localització dins la pàgina respecte un altre element o la seva situació zonal. També podem definir una sèrie de diccionaris on podem definir quines són les cadenes de caràcters que acceptaríem com a patró dins d'un camp concret. Aquests localitzadors són molt útils a la hora d'obtenir una aproximació vàlida de resultats provinents del l'OCR per cada camp, a més ens els ofereix ordenats per percentatge de validesa segons el que haguem configurat. A part de les opcions que ens ofereix els localitzadors configurables del KTM, també podem crear localitzadors definits des de scripts programats en WinWrap Basic.

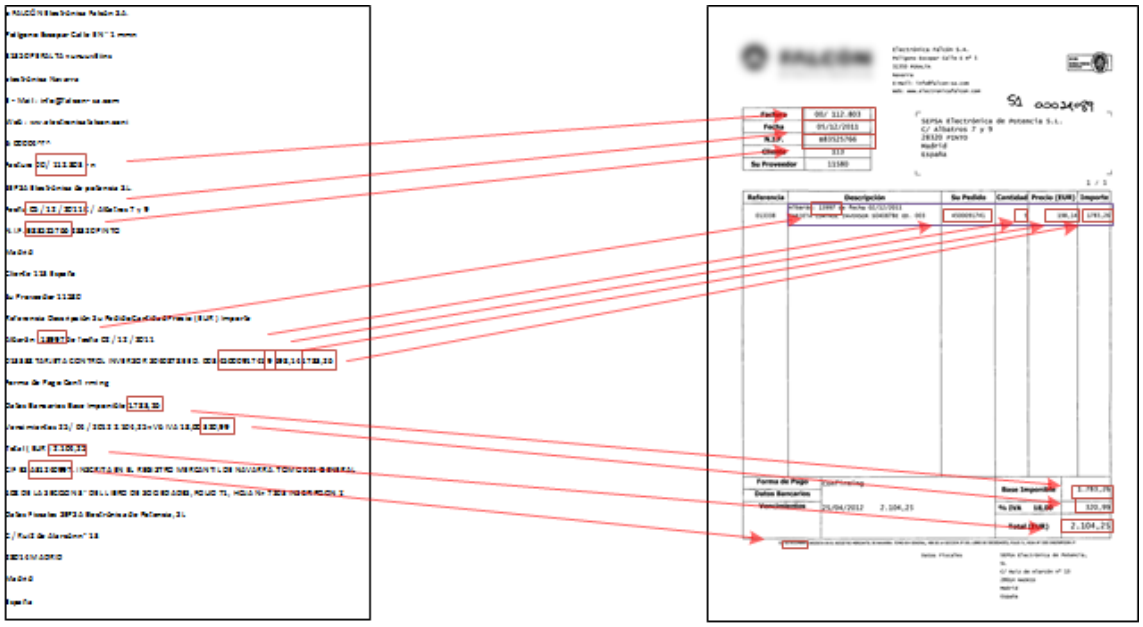


Figura 19. Procés d'extracció del document

El primer objectiu que volem complir a la hora de crear un projecte com aquest és el de poder aplicar aquesta solució a qualsevol tipus de factura que pugui arribar i que podem detectar la major part de camps sense necessitat de personalitzar la configuració i programació del projecte per aquell tipus de proveïdor. Per això, el primer pas és definir la localització de cada camp de manera genèrica, és a dir, de forma que es pugui aplicar al major nombre de factures possible amb un resultat vàlid. Més endavant podrem definir la localització pels camps d'aquells proveïdors que no hagin donat un bon resultat en les proves.

7.4.2.1 Cif del proveïdor

A continuació es mostra la configuració d'extracció del localitzador pel camp "cifprov":

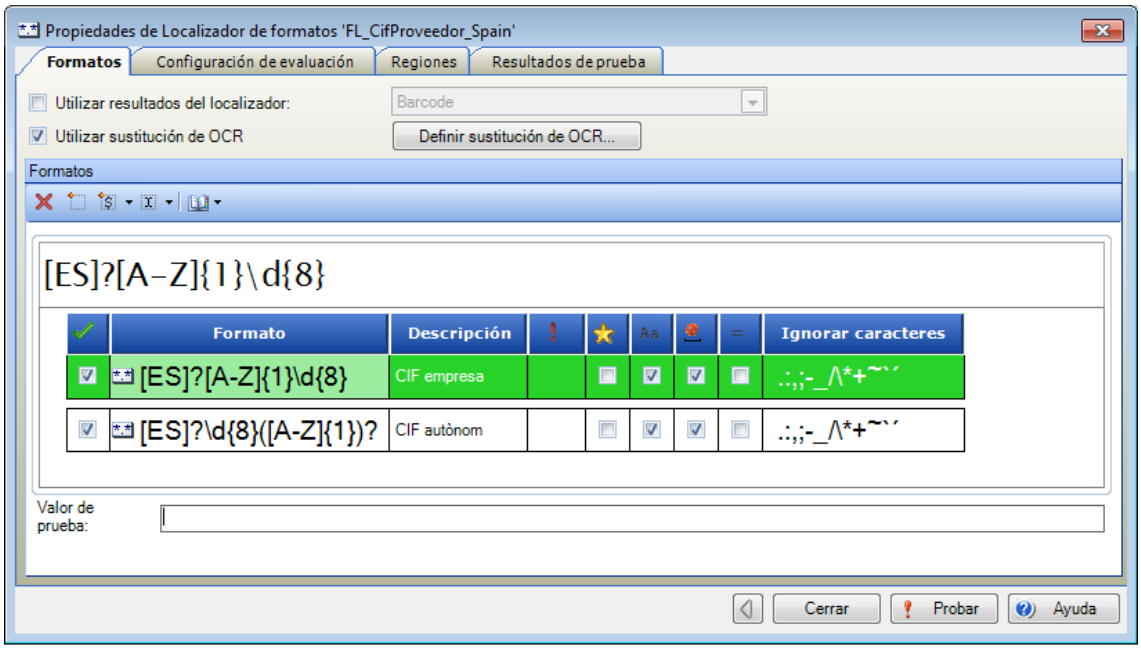


Figura 20. Configuració dels patrons d'extracció pel CIF del proveïdor

Aquí li estem indicant que el camp del CIF que ha de localitzar estarà compost amb un caràcter seguit de vuit xifres o vuit xifres seguides d'un caràcter (pel cas dels autònoms), podent haver-hi les sigles "ES" davant de la cadena. També podem veure com especifiquem els caràcters que ha d'ignorar per interpretar el patró, a la fase de formateig ja ens encarregarem de filtrar tots aquests caràcters que no ens interessa guardar.

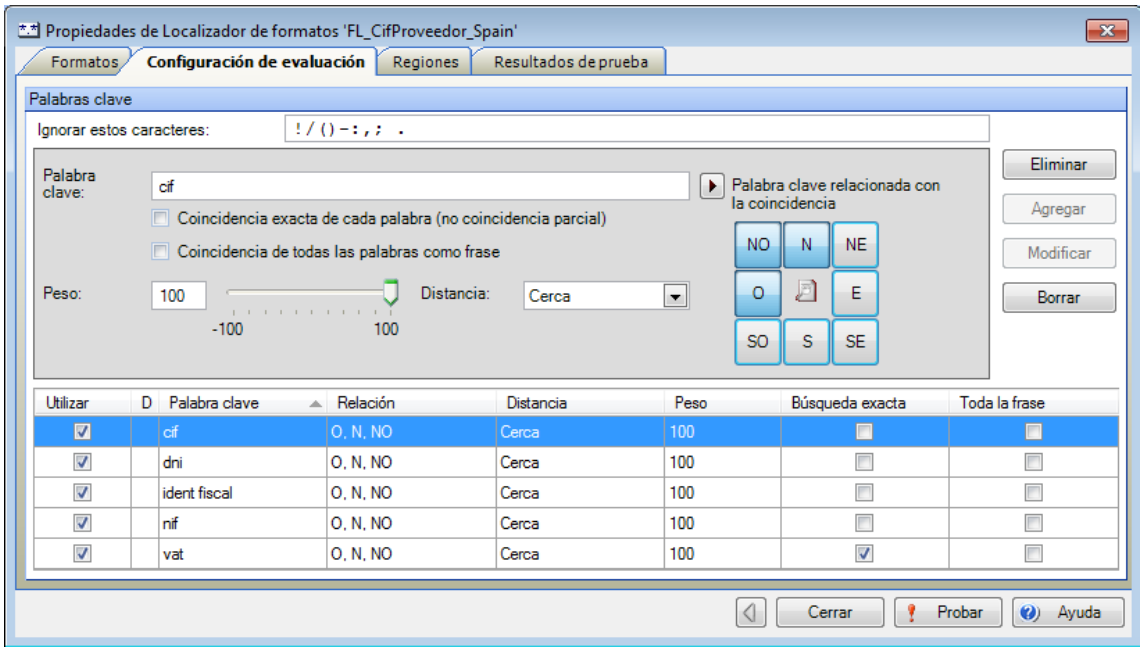


Figura 21. Configuració de les frases d'avaluació pel CIF del proveïdor

En aquesta figura podem definim les cadenes de caràcters que ens permeten identificar que el patró localitzat és un cif, ja que a prop s'ha trobat una de les paraules clau d'aquest camp. En aquest cas s'han definit *cif, dni, ident fiscal, nif* i *vat.*

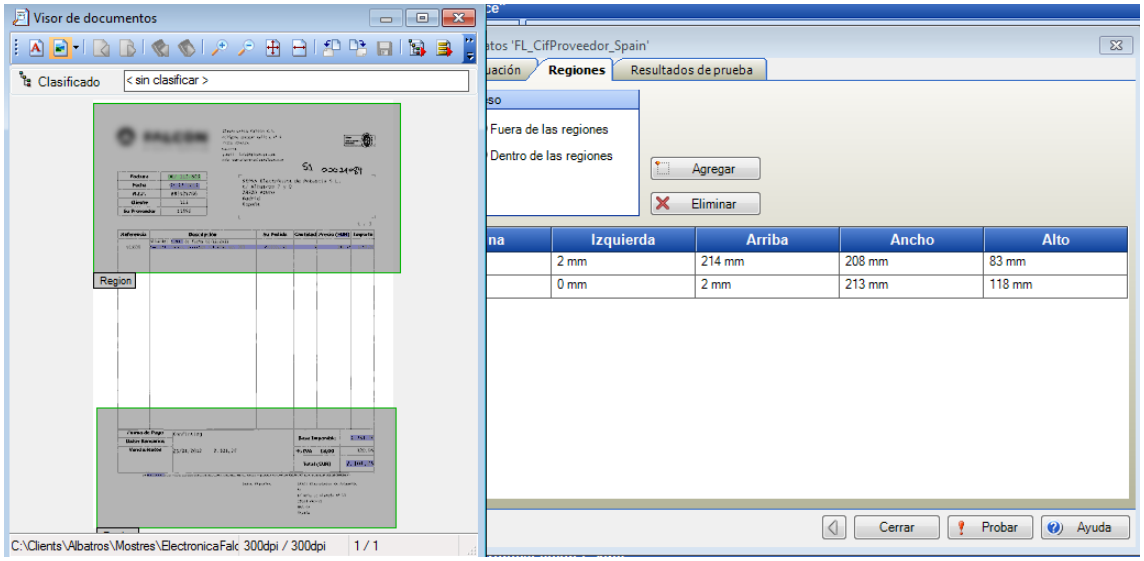


Figura 22. Configuració de les zones del document per realitzar l'extracció del CIF del proveïdor

A aquesta imatge es veu com li hem indicat les regions dins la pàgina on podrà trobar el CIF. La gran majoria de proveïdors situen el CIF a la part superior o a la part inferior de la factura sempre que aquest estigui de forma horitzontal. Pels casos en que el CIF es troba de manera vertical en el document s'ha desenvolupat un altre localitzador que és capaç d'identificar aquest tipus de situacions.

En el cas d'aquest camp tenim la problemàtica que a l'hora d'extreure un CIF el programa el confongui amb un CIF propi de la corporació, ja que aquest també sempre apareixerà a totes les factures. Per aquest motiu s'ha desenvolupat una funció que tracta el resultat obtingut del localitzador i només accepta aquelles cadenes de caràcters diferents als CIF propis de la corporació Albatros.

7.4.2.2 *Import total de la factura*

A continuació es mostra la configuració d'extracció del localitzador pel camp "totalfactura":

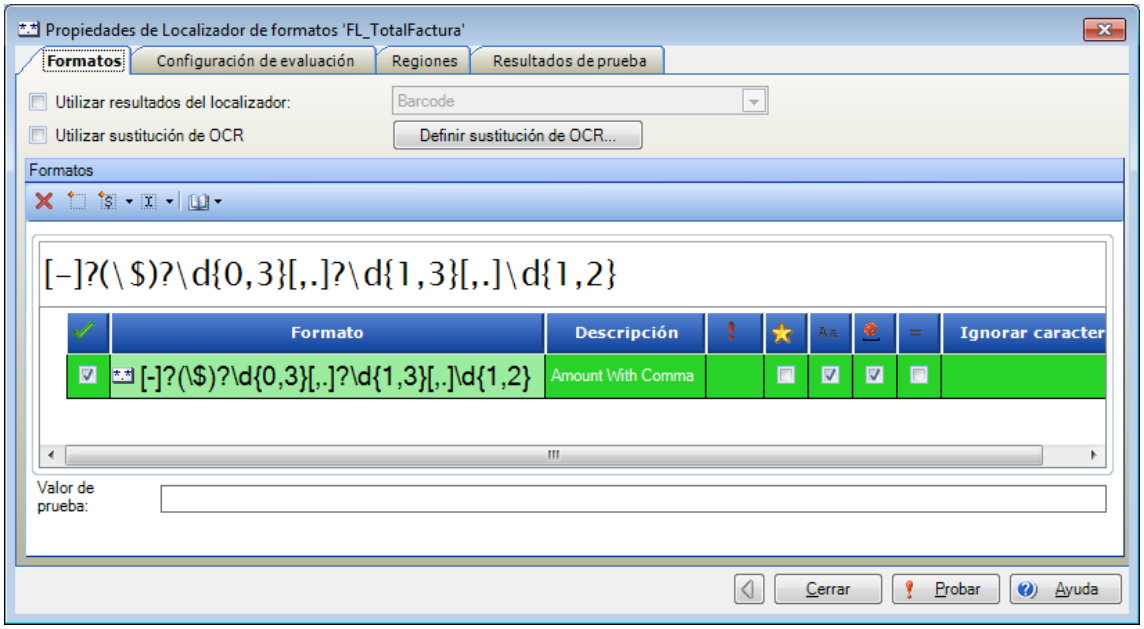


Figura 23. Configuració del patró d'extracció del total de la factura

En aquest cas, i en el de tots els imports, configurem el patró a trobar perquè pugui reconèixer qualsevol import.

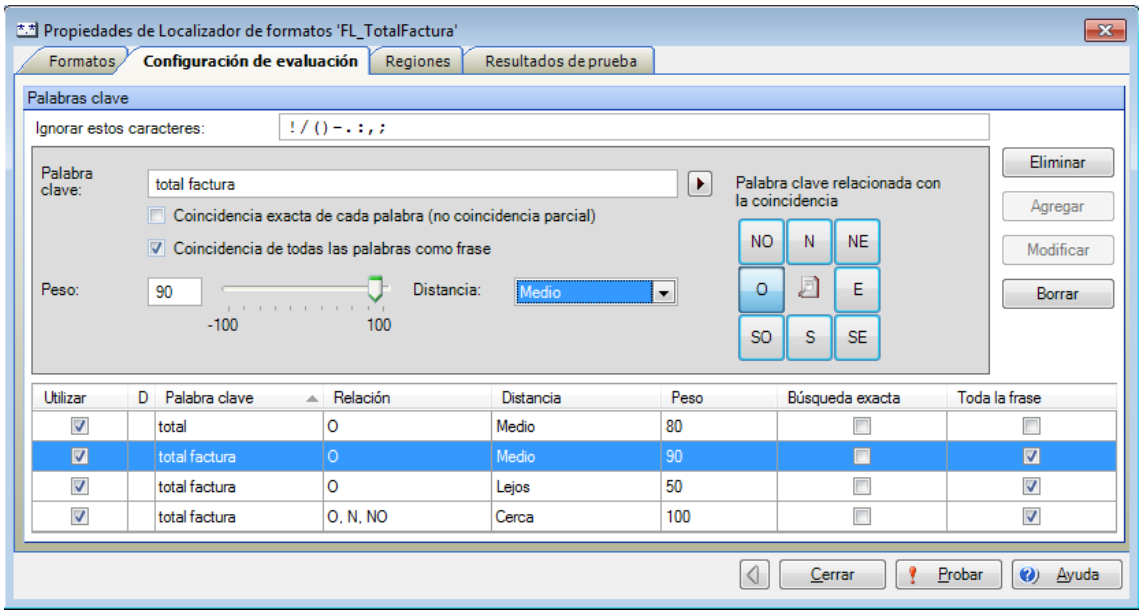


Figura 24. Configuració de les frases d'avaluació pel total de la factura

En aquesta imatge veiem les paraules clau definides pel total de factura. En aquest cas s'han definit diferents percentatges de validesa depenent de la distància a la que s'ha localitzat cada paraula clau. Aquests percentatges de validesa permetran un cop executat el localitzador determinar quin és el resultat amb més probabilitats de ser el total de la factura, evitant la confusió amb altres imports que puguin aparèixer al document.

Finalment, definim la regió del document on serà més probable trobar l'import total, que és lògicament al final de la pàgina.

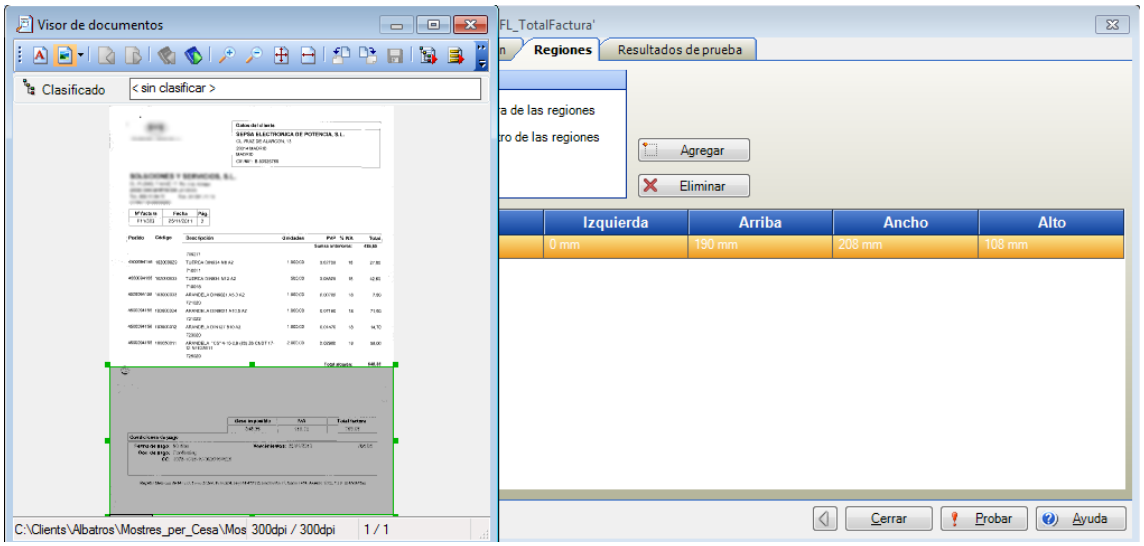


Figura 25. Configuració de les zones del document per realitzar l'extracció del total de la factura

7.4.2.3 Línies de comanda

Per les línies que contindran les factures amb línies de comanda s'ha configurat un localitzador basat en l'estructura de les línies i en els títols de capçalera que solen portar les taules de línies de comanda. Per fer-ho s'ha entrenat el localitzador identificant la capçalera de varies mostres de factures amb comanda. A més, per cada mostra s'ha assignat cada títol de la capçalera a entrenar a la columna de la taula anomenada *Detalle* que volem que li correspon. Aquest procés s'ha hagut de realitzar amb un conjunt de mostres variades i prou gran, ja que no sempre apareixen els mateixos títols per identificar cada columna. Per exemple, per la columna *dt_preciototal* corresponent a l'import total de la línia, el podem trobar identificat com "Importe total", "Total", "Precio total", etc.

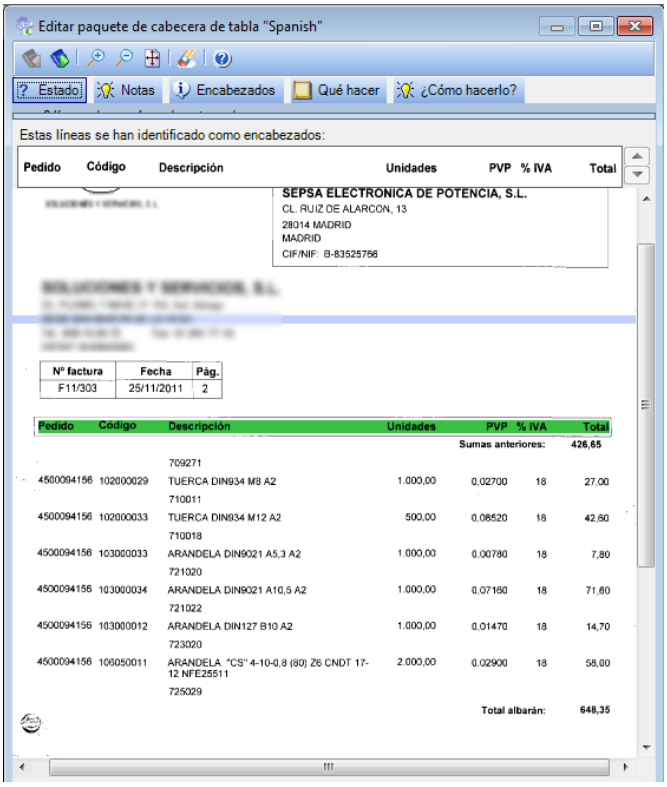


Figura 26. Configuració de l'entrenament per la detecció automàtica de les capçaleres de les línies de comanda

Si volem configurar la captura de línies per un proveïdor en concret podem sofisticar la localització definint un localitzador de línies manual. Al fer-ho es parteix d'una mostra d'una factura i es selecciona una línia assignant cada valor d'aquesta a la columna corresponent. A partir d'aquest moment el programa sempre intentarà buscar aquest format de línia per aquest proveïdor.

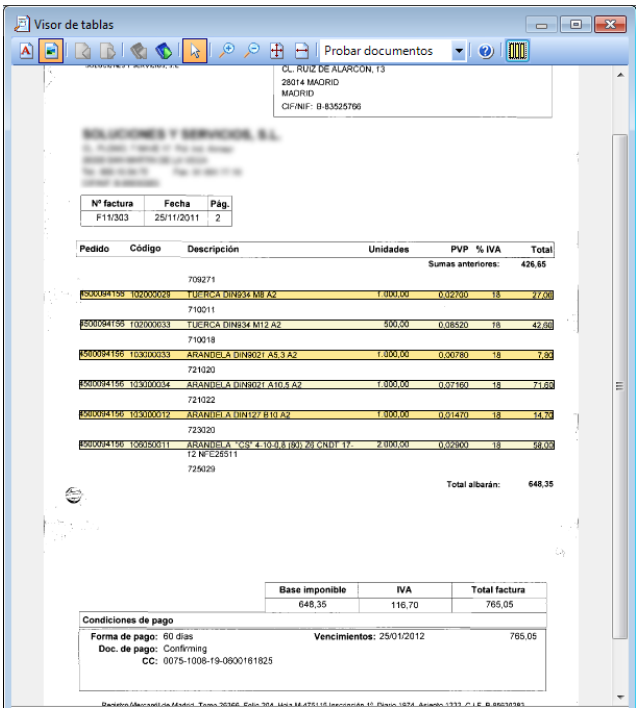


Figura 27. Configuració de la localització de línies per la plantilla d'un proveïdor

7.4.3 Formateig

Un cop superada l'etapa d'extracció i per tal d'aconseguir una validació òptima del format de cada camp obtingut es procedeix a desenvolupar el codi que s'aplicarà a tots aquells camps que tenen un format conegut. En aquesta etapa hem de definir quins són els caràcters que habilitem perquè siguin vàlids a cada camp. Amb això aconseguim evitar possibles signes de puntuació que s'hagin pogut captar de forma errònia a través de l'OCR i també que els valors que obtenim de cada camp siguin homogenis.

Amb aquest propòsit s'han desenvolupat varies funcions de formateig a través de les quals podrem filtrar caràcters no desitjats de les cadenes de caràcters. També s'ha definit el format de cada camp per poder-los guardar amb el mateix patró evitant així, problemes a l'hora de validar el contingut amb la base de dades.

Hi ha un grup de camps que tindran de manera evident el mateix tractament, aquests són els camps numèrics com els imports. Per aquest tipus de camps s'ha creat una funció que, tenint en compte els errors més comuns que comet l'OCR a la hora

d'interpretar els patrons numèrics, substitueix caràcters per números. Com podem veure són caràcters que fàcilment poden crear confusió a qualsevol algorisme que hagi de reconèixer els números en qüestió.

```
Public Function formatearNumeros(ByVal num As String) As String
   On Error GoTo Exception
      Dim numaux As String
      numaux = num
      numaux = Replace(numaux, "S", "5")
      numaux = Replace(numaux, "s", "5")
      numaux = Replace(numaux, "C", "0")
      numaux = Replace(numaux, "O", "0")
      numaux = Replace(numaux, "o", "0")
      numaux = Replace(numaux, "I", "1")
      numaux = Replace(numaux, "i", "1")
      numaux = Replace(numaux, "l", "1")
      numaux = Replace(numaux, "L", "1")
      numaux = Replace(numaux, "|", "1")
      numaux = Replace(numaux, "g", "9")
      numaux = Replace(numaux, "G", "6")
      numaux = Replace(numaux, "e", "6")
      formatearNumeros = numaux
      Exit Function
   Exception:
End Function
```

Figura 28. Fragment de codi referent a la substitució de caràcters que poden confondre l'OCR

Ara per a cada camp definim els caràcters que són vàlids i quin serà el format amb el que es guardarà el camp. A continuació, es mostren varis exemples del resultat de diferents camps un cop formatejades les cadenes de caràcters obtingudes de l'etapa d'extracció.

7.4.3.1 Cif del proveïdor

Després de filtrar caràcters no desitjats desenvolupem la lògica que és pròpia de cada camp. En el cas del CIF hem de tenir en compte que, amb els CIFs espanyols, a vegades apareixen els caràcters "ES" davant del codi. També es tenen en consideració altres possibles confusions tenint en compte el patró que s'està buscant, com pot ser la substitució del caràcter "B" pel caràcter numèric "8" en el cas de que aquest estigui davant de 8 xifres més.

Aquests podrien ser exemples del resultat de l'etapa de formateig pel camp del CIF:

8-12345 678	**->**	**B12345678**
ESB.123'45678	**->**	**B12345678**

7.4.3.2 Imports

En el cas dels imports filtrem tots els caràcters que no siguin numèrics i executem les substitucions per aquests tipus de camps. Si el camp està buit perquè no s'ha localitzat cap resultat se li assigna l'import per defecte que serà el "0,00".

Aquests podrien ser exemples del resultat de l'etapa de formateig per un camp d'import:

G8,O3	**->**	**68,03**
1,92I.S4	**->**	**1921,54**

7.4.3.3 Número de comanda

El valor del número de comanda té un format molt concret a la corporació Albatros, però pot aparèixer de maneres diferents depenent del proveïdor. Aquest número sempre comença per 2 xifres seguides de 3 zeros i 5 xifres més. Però sovint els 3 zeros són omesos als documents de les factures, així que per homogeneïtzar el format i posteriorment poder validar el valor contra bases de dades definim el format amb totes les xifres i afegim els zeros en el cas de que no apareguin.

s/45-00012345	**->**	**4500012345**
45\12345	**->**	**4500012345**

7.4.4 Validació

Dins de l'etapa de validació podem distingir dos grans grups: els camps que es validen contra dades conegudes i els camps que es validen a partir de regles de negoci i fórmules aritmètiques partint d'altres camps del mateix document.

7.4.4.1 Validació contra bases de dades

Normalment, les empreses mantenen bases de dades de les dades de proveïdors, de comandes i d'albarans. Com que són dades que hem de trobar a les factures i que han d'existir podem fer la validació d'aquestes contra bases de dades i assegurar-nos així, que les dades extretes d'aquests camps són correctes. Per fer-ho s'ha desenvolupat a partir de ADODB el codi necessari per establir una connexió ODBC (*Open Database Connectivity*) amb una base de dades SQL. ADODB és un conjunt de llibreries de bases de dades que permet als programadors desenvolupar aplicacions de manera portable mentre que ODBC és un estàndard d'accés a bases de dades desenvolupat per Microsoft. L'objectiu d'aquesta configuració és que ens permeti l'execució de qualsevol consulta des de l'aplicació sense que importi quin és el sistema gestor de bases de dades. D'aquesta manera podem executar consultes SQL dins del codi en WinWrap Basic per obtenir i comparar les dades desitjades.

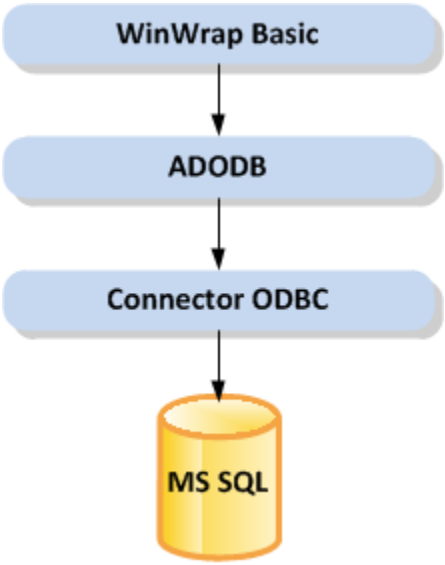


Figura 29. Esquema de connexió a la base de dades

A l'hora de triar un mètode de connexió també s'ha tingut en compte OLEDB, que suporta un nombre major de gestors de bases de dades. En el cas de la corporació Albatros, però, treballem sobre Microsoft SQL Server, plataforma desenvolupada per Microsoft, i s'ha comprovat que el rendiment de les llibreries de ADODB, al estar

basades en plataformes de Microsoft, és superior i funcionen de manera més eficient que OLEDB.

Els camps que es volen validar contra bases de dades de l'empresa client són el del CIF del proveïdor i el del número de comanda, és a dir, els camps *cifprov* i *numped*, respectivament. A més, pel número de comanda volem validar les línies de detall.

7.4.4.1.1 CIF del proveïdor

Pel cas del proveïdor, a més de validar el camp, també es vol obtenir el seu nom, el seu país i el codi SAP que té associat per introduir aquestes dades als camps corresponents. D'aquesta manera aconseguim una taxa d'errors molt inferior ja que aquests camps sempre seran vàlids i només haurem d'acorar l'obtenció del CIF per obtenir els demés camps.

En el cas que l'extracció no hagi sigut satisfactòria i el programa no hagi sigut capaç de trobar el CIF del proveïdor correctament s'ha configurat un cercador per a què l'usuari a través del formulari de validació pugui cercar el proveïdor en qüestió a partir del CIF, el número de SAP, el nom o la societat a la que pertany.

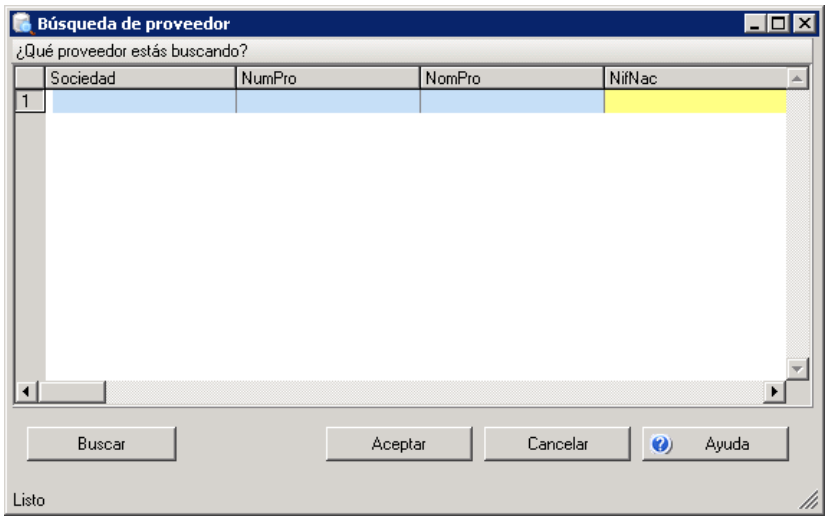


Figura 30. Cercador de proveïdors

7.4.4.1.2 Número de comanda

En el cas del número de comanda, el primer que comprovarem, de la mateixa manera que amb el CIF, és si el contingut dels camps existeix a la base de dades. A continuació, si s'ha trobat el valor es comprova si la comanda només té una línia, si és així, obtenim les demés dades de la línia, que són la posició de la línia, el número d'albarà, el codi d'article, la descripció, la quantitat, el preu unitari i el preu total, és a dir, els camps *dl_posicion, dl_numalb, dl_articulo, dl_descripcion, dl_cantidad, dl_preciounitario, dl_preciototal*, respectivament. D'aquesta manera aconseguim millorar el percentatge d'encerts en les factures que només tenen una línia, ja que no depenem de que l'extracció hagi pogut capturar correctament tots els valors del camps. A les factures de més d'una línia l'obtenció de camps es produeix a partir del número de comanda, del número d'albarà i del codi d'article.

Pels casos en que el valor d'alguna línia no s'hagi pogut obtenir o s'hagi obtingut de forma errònia s'ha desenvolupat un cercador perquè l'usuari tingui la possibilitat d'obtenir la les línies de la comanda desitjada. Aquest cercador consta de dos diàlegs modals desenvolupats a partir de les opcions que ens ofereix el llenguatge basat en WinWrap Basic.

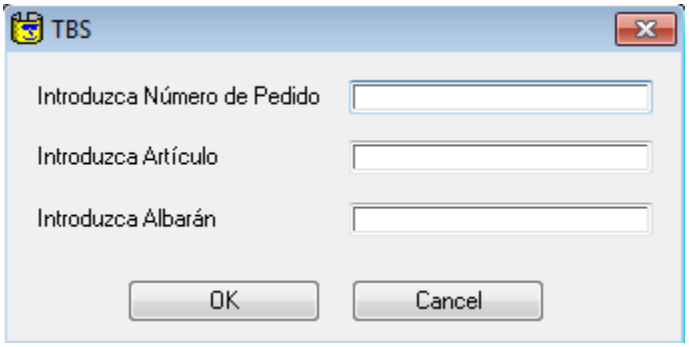


Figura 31. Cercador de línies de comanda

Aquest cercador és capaç de llistar totes les línies que existeixen a la base de dades indexades per número de comanda, per codi d'article i/o per número d'albarà. A més, a petició del client s'ha habilitat la cerca per més d'un número d'albarà. Aquesta cerca es realitzarà introduint més d'un número d'albarà separats per punts i comes.

D'aquesta manera es facilita la feina de l'usuari validador, ja que en moltes ocasions les línies de les comandes provenen de més d'un albarà.

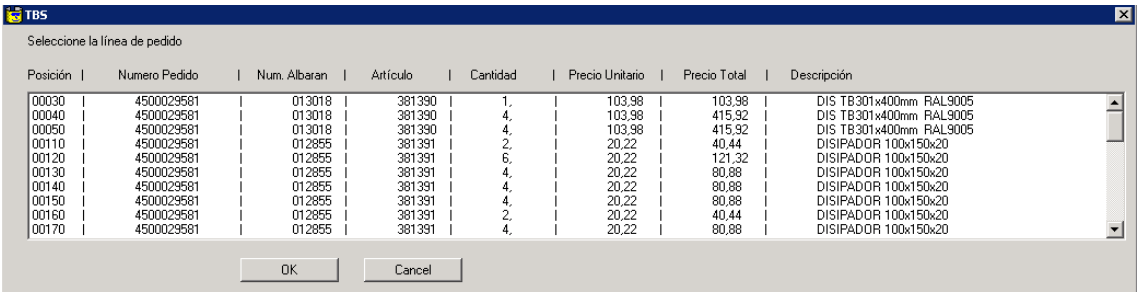


Figura 32. Resultat del cercador de línies de comanda

A la finestra on es llisten les línies resultants de la consulta l'usuari té la possibilitat de seleccionar una o més línies. Un cop seleccionades i polsat el botó "OK" aquestes línies s'afegeixen a la taula del formulari de validació assignant cada camp a la columna corresponent de la taula.

7.4.4.2 Validació amb regles de negoci

En aquesta part de la validació utilitzem la lògica pròpia dels documents en qüestió, en el nostre cas com el que volem és validar els valors dels camps que corresponen a factures, haurem de verificar que aquests valors són els correctes.

Els camps que podem validar seguint les regles de negoci corresponents a una factura estàndard són els imports:

- totalneto0
- pctimpuesto1
- impimpuesto1
- totalneto1
- pctimpuesto2
- impimpuesto2
- totalneto2
- pctimpuesto3
- impimpuesto3
- totalneto3
- pctimpuesto4
- impimpuesto4
- totalneto4
- pctirpf
- impirpf
- baseimp
- totalimpuestos
- totalfactura

Així, la fórmula que s'aplicaria per calcular l'import total de la factura és:

$$\textbf{totalfactura = totalneto0} + \sum_{k=1}^{4}(\mathbf{totalneto(k)} + \mathbf{impimpuesto(k)}) - \mathbf{impirpf}$$

També se'ns ha requerit camps de càlculs intermitjos i subtotals que responen a altres fórmules. Aquest és el cas de la base imposable i del total dels impostos:

$$\textbf{baseimp} = \sum_{k=0}^{4} \mathbf{totalneto(k)}$$

$$\textbf{totalimpuestos} = \sum_{k=1}^{4} \mathbf{impimpuesto(k)}$$

En el cas dels percentatges, s'ha tingut en consideració que al ser impostos, a aquests camps no s'aplicarà qualsevol valor, sinó que és caracteritzen per tenir un rang de valors possibles i determinat. Per això, per fer una validació més eficient, calculem els percentatges a partir dels imports i comparem els valors resultants amb els rang de valors possibles per comprovar que és un percentatge possible i aplicable en una factura. D'aquesta manera la fórmula amb la que calcularem els percentatges serà la següent:

$$\textbf{pctimpuesto(k) = impimpuesto(k) / totalneto(k) * 100}$$

$$\textbf{pctirpf = baseimp / impirpf * 100}$$

Aquestes validacions aritmètiques s'executen quan els camps implicats tenen imports vàlids, ja sigui per camps que s'hagin captat digitalment i hagin passat per l'etapa d'extracció o que s'hagin introduït directament des de teclat al formulari de validació. En el cas que no es compleixin alguna de les validacions es marcaran els camps implicats en vermell i es mostrarà un missatge d'error amb l'explicació de la causa que l'ha provocat.

Figura 33. Mostra de la validació dels imports en el formulari de validació

7.5 Formulari de validació

Tot i que s'aplica una validació automàtica al document hem de tenir en consideració que, encara que podem aconseguir un percentatge molt alt d'encerts mitjançant l'extracció, el formateig i la validació automàtica, no sempre podrà ser del 100%. L'algorisme OCR a vegades, ja sigui per la grandària excessivament petita del text o per algun tipus de confusió pot errar en el moment de l'extracció i, per això s'ha cregut necessari la creació d'un formulari de validació on l'usuari pugui confirmar o modificar les dades que s'han anat processant al llarg de tot el circuit.

S'ha de pensar que al ser documents de naturalesa fiscal és important que les dades que s'exportin siguin el màxim de correctes possibles, ja que si l'error es detecta

posteriorment sempre serà més costós de corregir. Un usuari del gestor documental encara podria modificar les dades de les factures a través de l'aplicació web, però aquests perfils d'usuaris estan més enfocats a l'aprovació de les factures, és a dir, a prendre decisions en funció del contingut d'aquestes i no tan a corregir errors. En el cas que l'error es detectés quan la factura ja està comptabilitzada dins de l'ERP s'hauria de crear un document rectificatiu i el procés es complica i s'encareix bastant més. Tot això ens porta a pensar que necessitem un formulari on un usuari pugui revisar i comprovar la correctesa de les dades.

A través d'aquest formulari l'usuari validador podrà validar totes les dades que s'han anat assignant i modificant a través de tot el procés comprovant si són correctes i, fins i tot, podent corregir possibles errors del document en paper. Per dissenyar aquest formulari s'ha desenvolupat una interfase de validació que s'adaptarà a les necessitats de l'usuari validador de manera que pugui veure clarament quins són els camps que s'han validat amb un percentatge de confiança suficientment alt perquè el programa asseguri la validesa de les dades que conté, i de quins no podem estar segurs i necessiten de revisió per part de l'usuari.

Aquesta interfase es divideix de forma vertical en dos panells. Al panell de la dreta es mostra la imatge de la factura que s'està validant, d'aquesta manera l'usuari pot comparar els resultats de la captura i els valors que figuren a la imatge de la factura. L'usuari pot interaccionar amb la imatge fent zoom, rotant-la i tenint la possibilitat de seleccionar el contingut per assignar-lo al camp que ell cregui convenient, ja que el resultat de l'execució OCR està totalment vinculat amb la imatge. Si es tracta d'una factura de més d'una pàgina es pot navegar entre elles amb facilitat, així com passar d'una factura a una altra.

Al panell de l'esquerra es mostra el formulari de validació amb tots els camps agrupats en funció del contingut. El grup superior conté el tipus de factura i dos botons per poder canviar canviar-lo, de factura amb comanda a factura sense comanda, o viceversa (MM i FI). També es mostra si la el document és una factura o un abonament i la societat seleccionada. A continuació, es mostra la capçalera de la factura que engloba el número de factura, la data de factura, el CIF del proveïdor amb totes les

dades associades i la moneda. Mitjançant les tecles F6 i F7 podrem accedir als cercadors de proveïdors i comandes, respectivament.

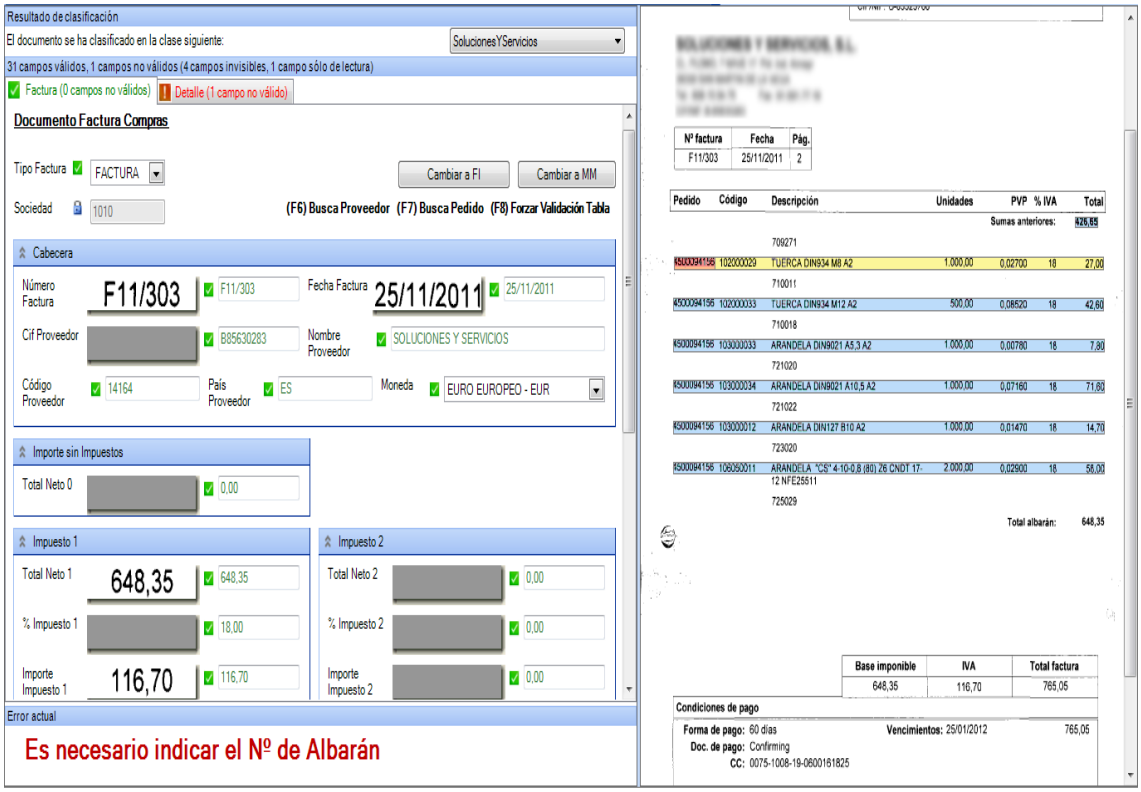


Figura 34. Formulari de validació (part superior)

A la part inferior es mostren tot els grups d'imports agrupats en 4 grups d'impostos (una factura tindrà com a molt 4 tipus d'impostos) i, finalment, els imports totals, incloent l'impost d'IRPF i les despeses de transport.

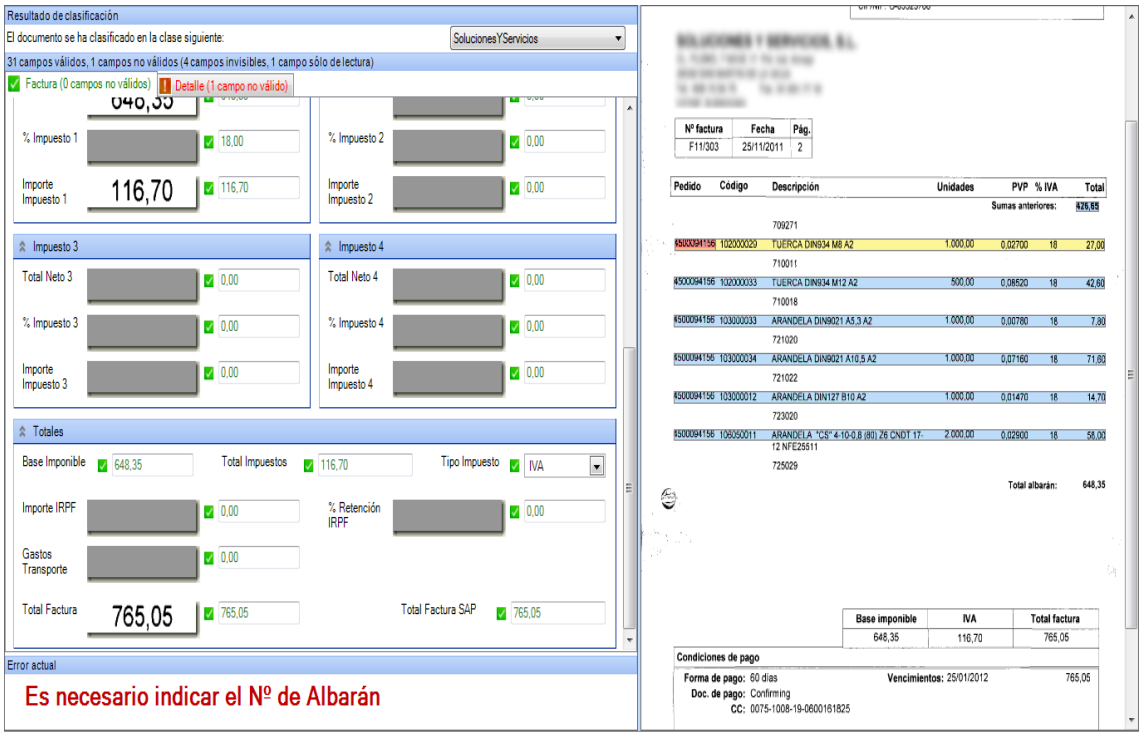


Figura 35. Formulari de validació (part inferior)

En el cas que sigui una factura amb comanda (MM) es mostra una segona pestanya que conté la informació referent a les línies de la comanda. En el cas que l'usuari vulgui validar les dades de la taula sense disposar de totes les dades de les línies té la possibilitat de forçar la validació (posar en verd) de tots els camps de la taula per poder continuar amb el processament de la factura en qüestió (F8) .

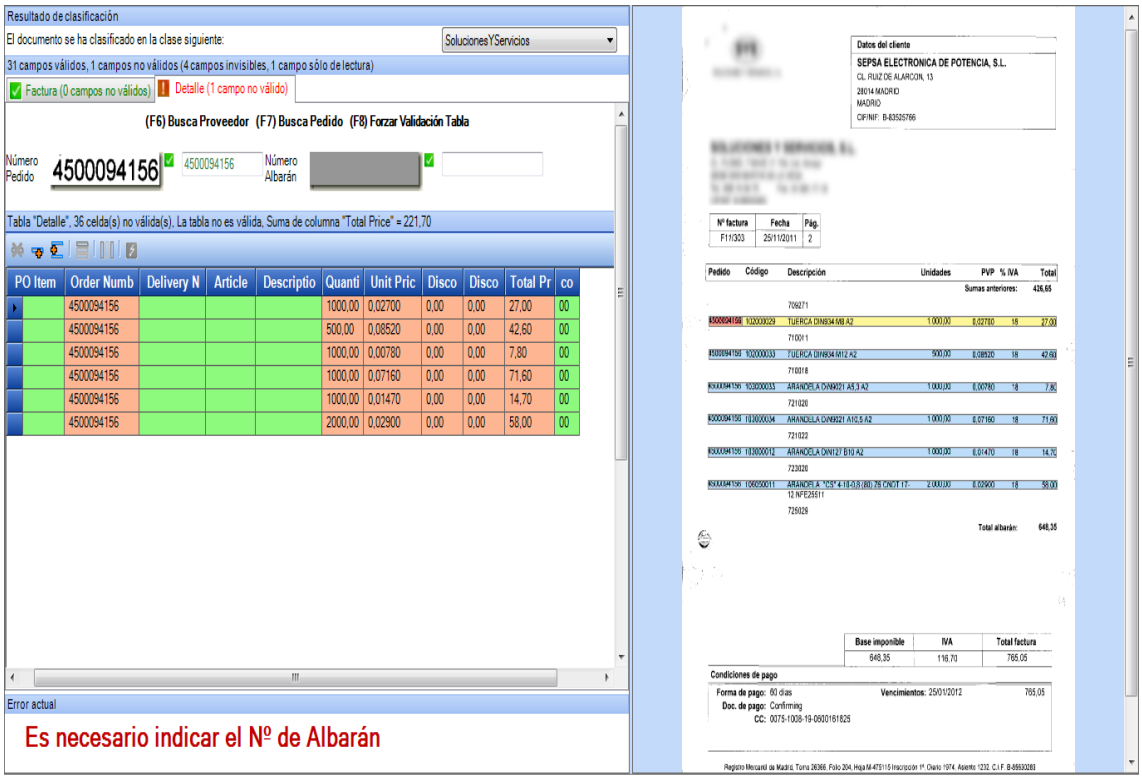


Figura 36. Formulari de validació (pestanya de les línies de comanda)

Com que l'objectiu d'aquesta etapa és que l'usuari que valida la factura pugui observar de manera ràpida i fàcil si cada camp és correcte, es mostren només en color verd els camps que s'han extret amb una confiança de més del 80%. Aquesta confiança és el percentatge d'encert entre la cadena de caràcters que obtenim de l'algorisme de l'OCR per aquell camp combinat amb el patró configurat pel localitzador d'extracció.

El formulari el creem a partir de la interfície de KTM. Mitjançant codi en Win Wrap Basic es determina el comportament de les interaccions que pot fer l'usuari, tals com les de polsar un botó amb el ratolí, validar un camp en concret o tot el document, tancar una finestra, etc.

7.6 Exportació a la base de dades

A través de Kofax Capture definim la correlació de les dades que obtenim de KTM amb els camps de la taula SQL. A part de les dades dels camps capturats i validats tenim la possibilitat d'emmagatzemar metadades com la data de creació del lot o el nom d'usuari que ha validat el lot, que ajudaran a poder dur un control més exhaustiu del

procés de captura. Així que, als camps que ja teníem definits, afegim les següents metadades:

Nom del camp	**Descripció**
UsuarioCaptura	Nom de l'usuari que ha creat el lot
FechaCaptura	Data de la creació del lot
IdEstacion	Equip des d'on s'ha validat el lot
NombreLote	Nom que se li ha donat al lot
NombreClaseLote	Nom de la classe del lot (d'una o vàries pàgines)

Taula 6. Llista dels camps referents a les metadades

A més, el mateix connector d'exportació també permet definir una taula dins la base de dades on guardarem la ruta de cada imatge processada. Cada línia d'aquesta taula tindrà un identificador que la correlacionarà amb les dades i metadades del document de l'altre taula. D'aquesta manera des del gestor documental podrem disposar de tota la informació referent al document a partir d'aquestes dues taules.

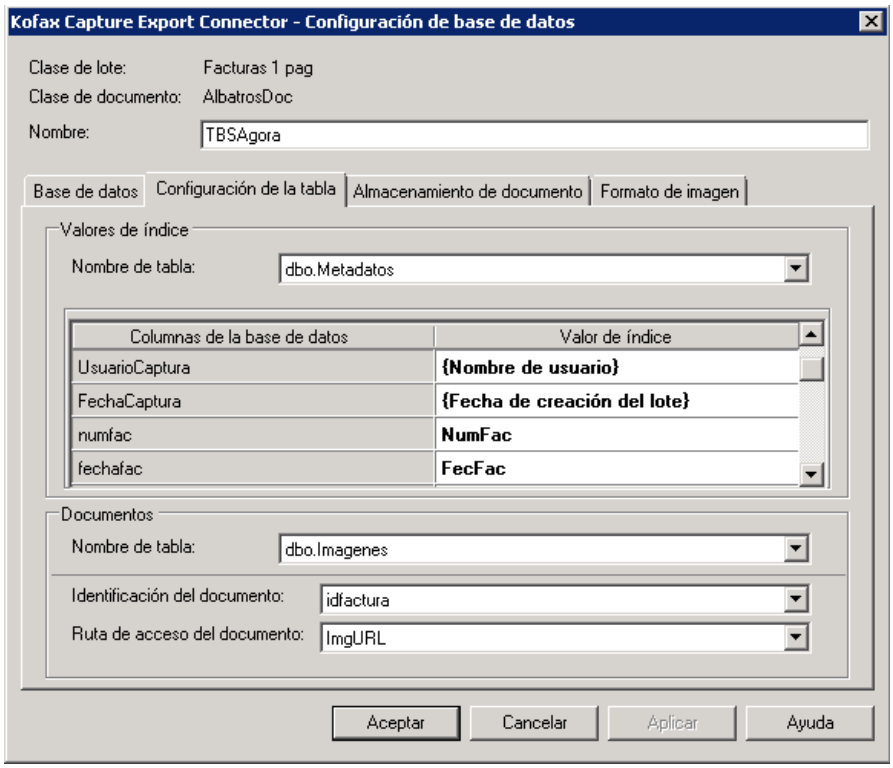


Figura 37. Finestra de configuració del connector de Kofax amb la base de dades

7.7 Integració amb el gestor documental

Per tal que les dades resultants del sistema de captura estiguin disponibles pel gestor documental necessitem un connector que extregui aquestes dades de la base de dades i les insereixi a l'aplicació web de manera ordenada i amb coherència amb la lògica de l'aplicació. TBSAgora es una aplicació que funciona sobre la plataforma Microsoft Sharepoint i que està desenvolupada en *C#*. Microsoft Sharepoint té una estructura molt concreta a l'hora de treballar amb les dades i els processos, és per aquest motiu pel que s'ha hagut de desenvolupar aquest mòdul d'integració.

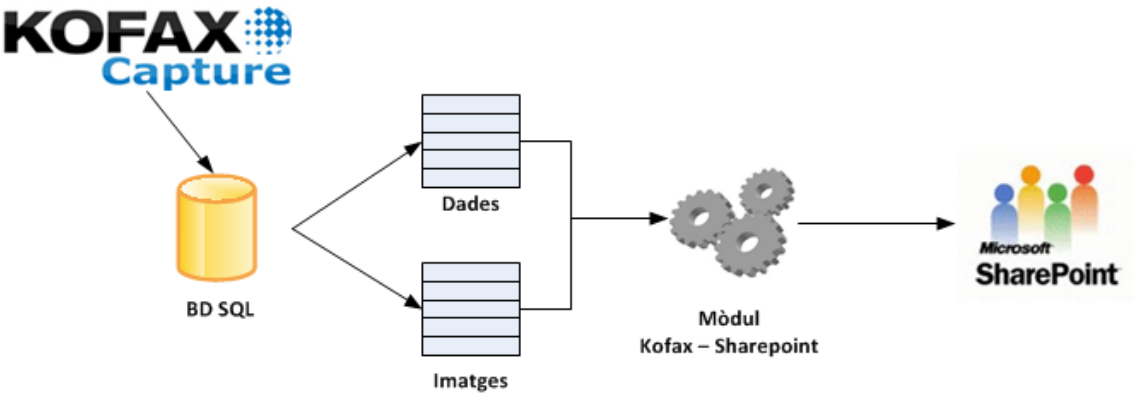


Figura 38. Esquema de funcionament del connector de Kofax amb Sharepoint

A Sharepoint s'han creat 3 llistes: una que contindrà els camps de les factures que no siguin línies, una altra amb les línies de les factures i una amb les imatges. En tots tres casos els documents estan identificats amb el valor únic incremental que proporciona Kofax i que relaciona la taula de dades i la d'imatges. Llavors quan s'executi aquest mòdul s'assignarà cada camp al que li correspongui dins de cada llista, tenint en compte si són línies o no. Les dades dels camps corresponents a les línies de comanda s'emmagatzemen a la base de dades separats per ";", per tant, s'ha de tenir en compte a l'hora d'assignar-les a cada llista de Sharepoint.

CampoKofax	CampoSharepoint	Formato	Descripcion	Linea
numfac	numfac	Texto	Nº de factura	No
fechafac	fechafac	Texto	Fecha de factura	No
numped	numped	Texto	Nº de pedido	No
numalb	numalb	Texto	Nº de albarán	No
codprov	codprov	Texto	Código SAP del proveedor	No
cifprov	cifprov	Texto	CIF del proveedor	No
nomprov	nomprov	Texto	Nombre del proveedor	No
tipofacsap	tipofacsap	Texto	Tipo de factura SAP: MM o FI	No
tipofac	tipofac	Texto	Factura o Abono	No
totalneto0	totalneto0	Numero	Importe exento de impuestos	No
pctimpuesto1	pctimpuesto1	Texto	Porcentaje del impuesto 1	No
impimpuesto1	impimpuesto1	Numero	Importe del impuesto 1	No
totalneto1	totalneto1	Numero	Importe total al que se haya aplicado el impuesto 1	No
pctimpuesto2	pctimpuesto2	Texto	Porcentaje del impuesto 2	No
impimpuesto2	impimpuesto2	Numero	Importe del impuesto 2	No
totalneto2	totalneto2	Numero	Importe total al que se haya aplicado el impuesto 2	No
pctimpuesto3	pctimpuesto3	Texto	Porcentaje del impuesto 3	No
impimpuesto3	impimpuesto3	Numero	Importe de impuesto 3	No
totalneto3	totalneto3	Numero	Importe total al que se haya aplicado el impuesto 3	No
pctimpuesto4	pctimpuesto4	Texto	Porcentaje del impuesto 4	No
impimpuesto4	impimpuesto4	Numero	Importe del impuesto 4	No
totalneto4	totalneto4	Numero	Importe total al que se haya aplicado el impuesto 4	No
pctirpf	pctirpf	Texto	Porcentaje de IRPF	No
impirpf	impirpf	Numero	Importe de IRPF a aplicar	No
baseimp	baseimp	Numero	Base imponible total de la factura	No
totalimpuestos	totalimpuestos	Numero	Suma total de los impuestos de la factura	No
totalfactura	totalfactura	Numero	Importe total de la factura	No
moneda	moneda	Texto	Moneda de la factura	No
dt_posicion	dt_posicion	Texto	Posición de pedido por línea de factura	Si
dt_numped	dt_numped	Texto	Nº de pedido por línea de factura	Si
dt_numalb	dt_numlab	Texto	Nº de albarán por línea de factura	Si

Figura 39. Llista dels camps a Sharepoint

Els fitxers d'imatge de les factures es serialitzen perquè estiguin disponibles dins la llista Sharepoint de l'aplicació web. A més, també és de rellevància comentar que es realitza un control de factures duplicades, que no permet poder entrar al gestor documental factures que ja s'havien processat anteriorment.

Mitjançant el camp anomenat "codtrans" es controlarà l'estat de la factura:

- **Valor "00":** Pendent d'entrar a Sharepoint
- **Valor "01":** Entrada a Sharepoint amb èxit
- **Valor "10":** No s'ha pogut entrar a Sharepoint perquè està duplicada
- **Valor "11":** S'ha produït un error al intentar entrar el document (fallo de connexió, etc.)

Aquest mòdul s'executarà cada 5 minuts com a servei de Windows, d'aquesta manera la tasca quedarà automatitzada. Quan s'executi totes les factures de la base de dades que tinguin el valor de "codtrans" a "00" s'intentaran entrar a les llistes de Sharepoint.

8. Digitalització certificada

A Espanya, la Digitalització Certificada de factures és el procés que permet obtenir còpies digitals de les factures amb valor original (el que fa possible destruir la factura en paper). En aquest projecte es va proposar al client una solució per dur a terme la digitalització certificada de les factures integrada amb la resta del sistema. Aquesta és una solució molt interessant, ja que al poder eliminar el document en paper el procés documental és pot optimitzar i economitzar de forma significativa.

Es tracta d'una tecnologia en un marc legal encara molt jove. Des de 2007, i segons la normativa *Agència Espanyola d'Administració Tributària* (AEAT), és possible digitalitzar els documents en paper d'interès tributari (factures, abonaments, etc.) i electrònica que doti els documents electrònics del mateix valor legal que els originals en paper. Els equivalents obtinguts, han custodiar-se en una base de dades documental que permetrà l'accés als mateixos i la seva aportació en cas d'auditoria o inspecció tributària.

Així doncs, no és necessària la custòdia dels documents en paper i es pot procedir a la seva eliminació. El procés implica la realització de signatures electròniques qualificades o reconegudes, l'intercanvi de certificats electrònics i el segellat de temps o *timestamping*[viii]. Aquesta tecnologia permet validar si un document ha estat modificat des que va realitzar el procés, i ha d'estar degudament homologada per l'AEAT.

8.1 Requisits

La normativa de l'AEAT es pot resumir en les següents disposicions:

1. S'han d'emprar formats estàndards d'ús comú: ISO 19005 (PDF/A)[ix], TIFF 6.0[x] o superior, JPEG2000[xi], Acrobat 5 (PDF 1.4[xii]) o superior i PNG[xiii]. En tots els casos utilitzant compressió sense pèrdues.
2. Nivell de resolució mínim dels documents digitalitzats definit a 200ppp[xiv].
3. Garantia d'imatge fidel i íntegra.
4. Per cada document digitalitzat, la base de dades documental ha d'incloure les imatges dels documents acompanyades d'un registre de dades amb tots els camps exigibles en la gestió dels llibres de registres.
5. Signatura de la base de dades mitjançant empremta o valor resum, signatura electrònica o signatura electrònica avançada.
6. Accés complet i sense demora als documents digitalitzats i els registres de dades.
7. Documentació que acrediti el compliment dels requisits exigits per l'Ordre EHA 962/2007[xv].
8. Optimització automàtica de la imatge per garantir la seva llegibilitat, de manera que tot el contingut del document original es pugui apreciar i sigui vàlid per a la seva gestió.

8.2 Signatura electrònica

Existeixen 3 tipus de signatura electrònica en funció del nivell de seguretat i integritat de cadascuna d'elles. S'han estudiat els 3 tipus per poder valorar quina és la opció més adequada en el nostre context.

8.2.1 Signatura electrònica simple

Adjunta dades en forma electrònica que poden ser usades per identificar l'autor de la signatura dels documents.

8.2.2 Signatura electrònica avançada

Aquests tipus de firmes permeten, a més de la identificació del signant, assegurar la integritat dels documents i la clau usada que van ser vinculats al signant. Per fer-ho s'usa la tecnologia PKI explicada més endavant.

8.2.3 Signatura electrònica reconeguda

En el procés de la digitalització certificada és necessari l'ús de la signatura electrònica reconeguda, consistent en un conjunt de dades o resum xifrat associat a un document que permet garantir la identitat del signant i la integritat del document, assegurant així que aquest no ha estat modificat. Perquè una signatura electrònica reconeguda sigui vàlida ha d'assegurar la integritat del document, i possibilitar així la detecció de possibles modificacions no autoritzades en el document. A més, ha de certificar l'autenticitat de l'emissor del document i registrar l'hora en què es va signar. Aquest procés es realitza mitjançant *timestamping*, que afegeix un segell de temps calculat a partir del hash del document i signat per una Autoritat de Segellat de Temps o TSA.

S'usa la tècnica criptogràfica anomenada Infraestructura de Clau Pública o PKI, a més d'una funció de hash, que s'encarrega de la encriptació del document.

La signatura digital o electrònica de la PKI fa ús de claus públiques i privades, de manera que l'emissor xifra un document amb la seva clau privada, clau que només ell coneix, i que només pot ser desxifrada amb la clau pública que té el receptor. Les relacions que identifiquen la clau pública amb el seu propietari, es denominen Certificats Digitals, i només poden ser emesos per una autoritat de certificació, que és una entitat especialitzada en la gestió d'aquests tipus de certificats, i que a més és de confiança per ambdues parts.

El procés funciona de la següent manera:

1. L'emissor crea un resum de les dades que vol enviar mitjançant una funció de hash creant així un codi hash o empremta digital que les representarà de manera unívoca.

2. L'emissor xifra el codi hash amb la clau privada que només ell coneix, obtenint així la signatura digital.
3. L'emissor adjunta la signatura digital a les dades, de manera que queden "signades".
4. L'emissor envia les dades signades al receptor.
5. L'emissor rep les dades signades i desxifra la signatura digital dels mateixos amb la clau pública de l'emissor, obtenint així el codi hash de les dades originals.
6. El receptor realitza una funció hash sobre les dades rebudes i comprova que siguin les mateixes que les obtingudes en el pas anterior. Si no és així, implicaria que les dades que ha rebut han estat modificades i no es corresponen amb les dades originals.

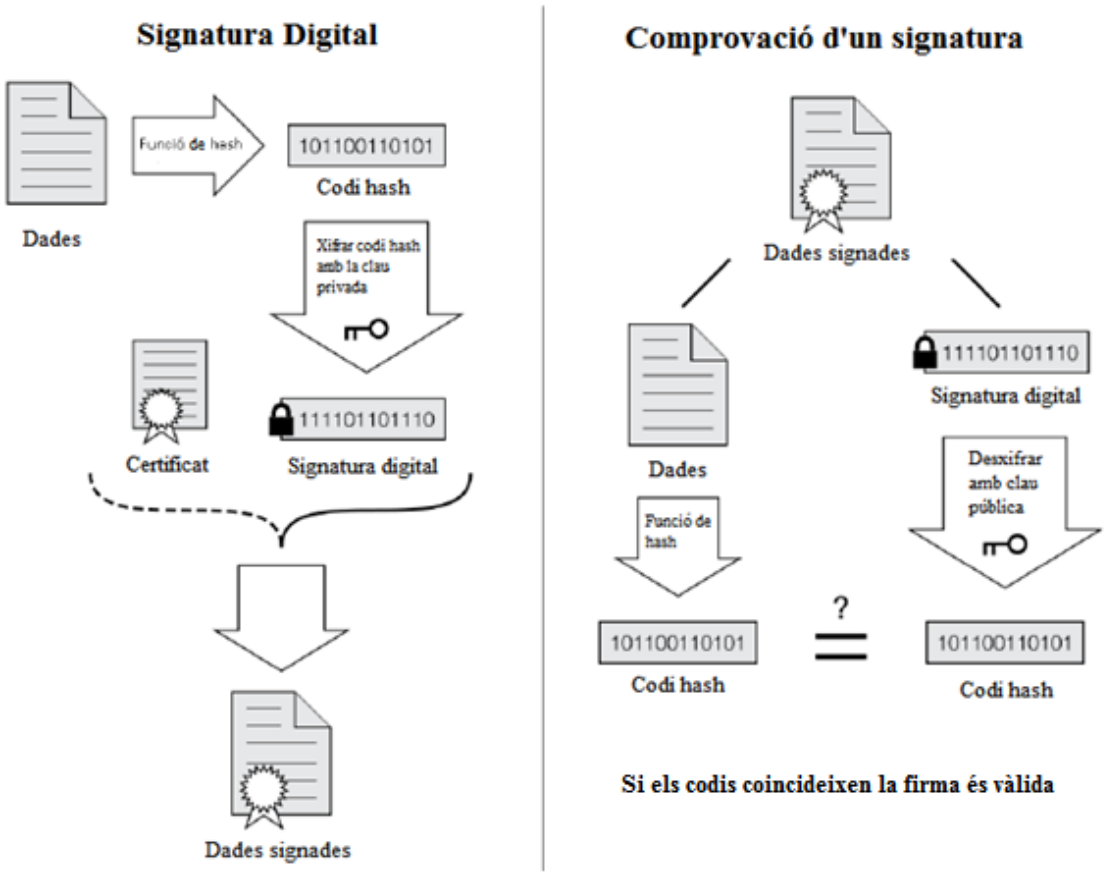


Figura 40. Etapes del procés i de la comprovació de la signatura digital

El software elegit per dur a terme aquest procés ha sigut *Legal Snap Scan* de *ANF AC & Tradise*. La principal raó per triar-lo ha sigut el seu disseny modular, que s'adapta perfectament a l'arquitectura d'aquest projecte, al contrari d'altres softwares que consten d'una solució completa i tancada.

A més, *Legal Snap Scan* té una gran experiència en el sector, ha estat el primer programari a ser homologat per l'AEAT i en complir amb l'Ordre EHA/962/2007 per la qual es desenvolupen determinades disposicions sobre facturació telemàtica i conservació electrònica de factures.

8.3 Integració amb Kofax

Per poder signar les factures que digitalitzem amb la solució presentada necessitem poder integrar el software de la signatura digital amb el software de Kofax. Per fer-ho s'ha desenvolupat un mòdul a partir de la API de Kofax perquè es pugui configurar dins del circuit de la solució com una fase més, però respectant tots els principis requerits per l'AEAT.

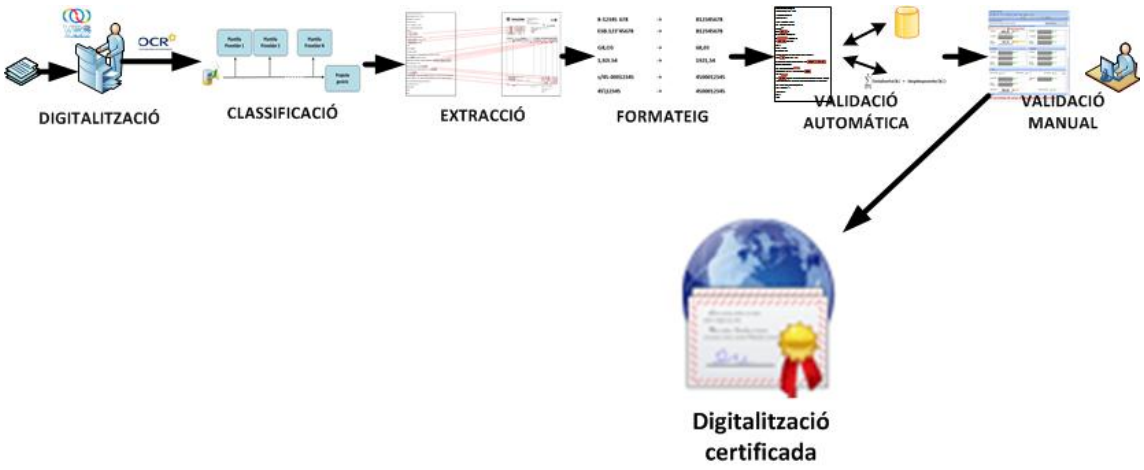


Figura 41. Circuit de la solució amb la digitalització certificada

S'ha decidit situar la certificació de la factura després de la validació manual perquè és en aquesta validació on l'usuari té la opció de revisar la factura i rebutjar-la en el cas que el document no tingui les dades necessàries o aquestes siguin errònies. En un inici la legislació obligava a signar la factura immediatament després d'escanejar-la, però en els últims anys s'ha flexibilitzat per tenir en compte aquest tipus de situacions.

Dins del nou mòdul que afegim a Kofax el procés de certificació s'implementa a través del software *Legal Snap Scan* i es divideix en tres fases que s'executaran seqüencialment:

- Generar signatura: Mitjançant el procés descrit a l'apartat anterior és genera la signatura electrònica reconeguda amb el certificat de l'entitat certificadora de *Tradise*.
- Agrupar les pàgines en documents: Quan generem els fitxers d'imatge en format TIFF des de Kofax obtenim un fitxer per cada pàgina del document, per tant, el que hem de fer és agrupar aquestes pàgines per poder tenir un arxiu per factura. Aquesta acció es realitza en una carpeta temporal. S'ha de tenir en compte que a la configuració de digitalització de Kofax hem aplicat les condicions de qualitat i resolució necessàries perquè la imatge obtinguda compleixi amb els requisits de l'AEAT. En aquest cas s'ha determinat una resolució de 300 ppp en format TIFF 6.0.

- Signar el document: És en aquest moment que ens connectem al servidor de *Tradise* per complir amb els requisits de *timestamping* marcant el document amb un segell de temps. En aquest moment si el procés ha funcionat correctament es genera un fitxer SLC que contindrà la informació del document signat i tot l'historial de passos que ha seguit el fitxer. Aquest fitxer es guarda a una altra carpeta que serà la que podrà ser auditada en cas de fer comprovacions o inspeccions de caire fiscal.

9. Gestor documental TBS Agora

A continuació, es descriu el funcionament del gestor documental TBS Agora explicant algunes de les possibilitats que tenen els usuaris una vegada les factures ja han estan processades per la plataforma de captura digital i estan en forma de dades a les llistes de Sharepoint dins l'aplicació web.

La pantalla inicial que veu l'usuari que ha d'aprovar les factures dins del gestor documental és un panell d'informació de les factures pendents i l'estat d'aquestes. També consta d'un cercador per poder localitzar i visualitzar les dades d'un document en concret.

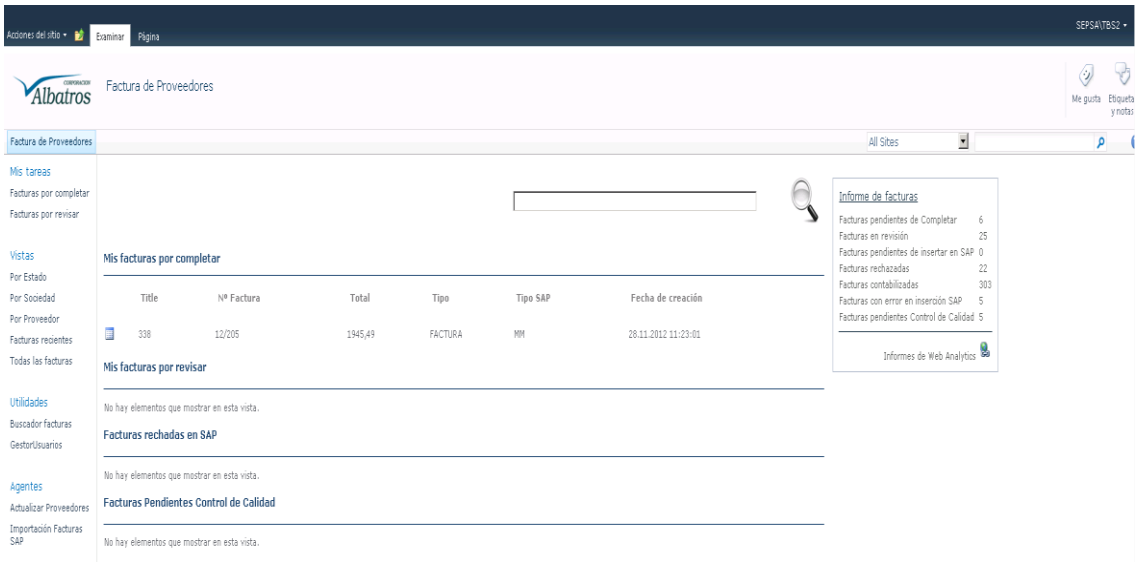


Figura 42. Pantalla inicial de TBS Agora

Si es desitja també es poden mostrar totes les factures i tots els camps guardats amb diverses vistes on es poden filtrar els documents per estat, societat o proveïdor.

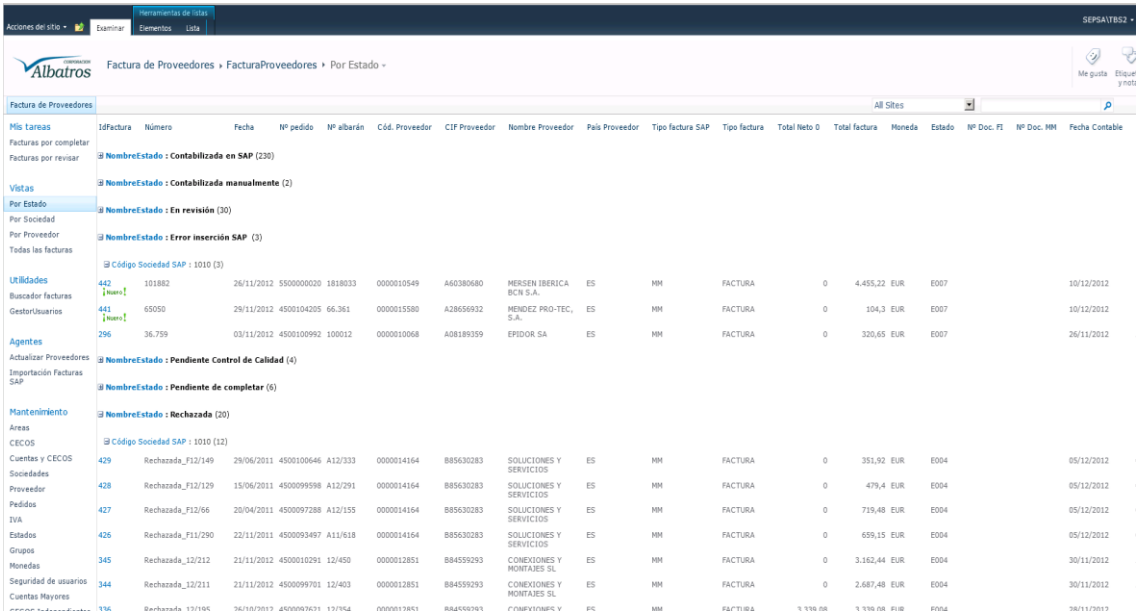


Figura 43. Vista de totes les factures per estat a TBS Agora

Quan es selecciona una factura l'usuari pot modificar totes les dades associades a aquesta. L'objectiu és que es puguin corregir errors que s'hagin pogut obviar en la fase de validació dins la plataforma de captura digital. També es podrà afegir informació d'índole fiscal, necessària per poder comptabilitzar la factura a SAP. Informació com les condicions o les vies de pagament, que són dades que no apareixen al propi document i que ha de complimentar l'usuari.

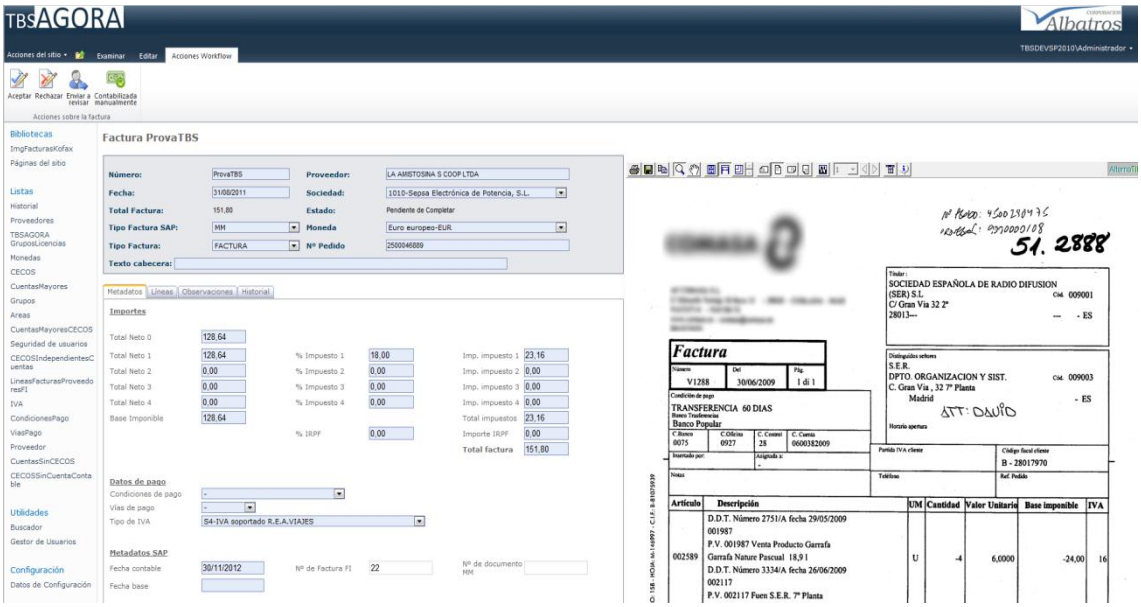


Figura 44. Pantalla de visualització de les dades d'una factura a TBS Agora

A la següent figura veiem com es visualitzen les línies de la factura.

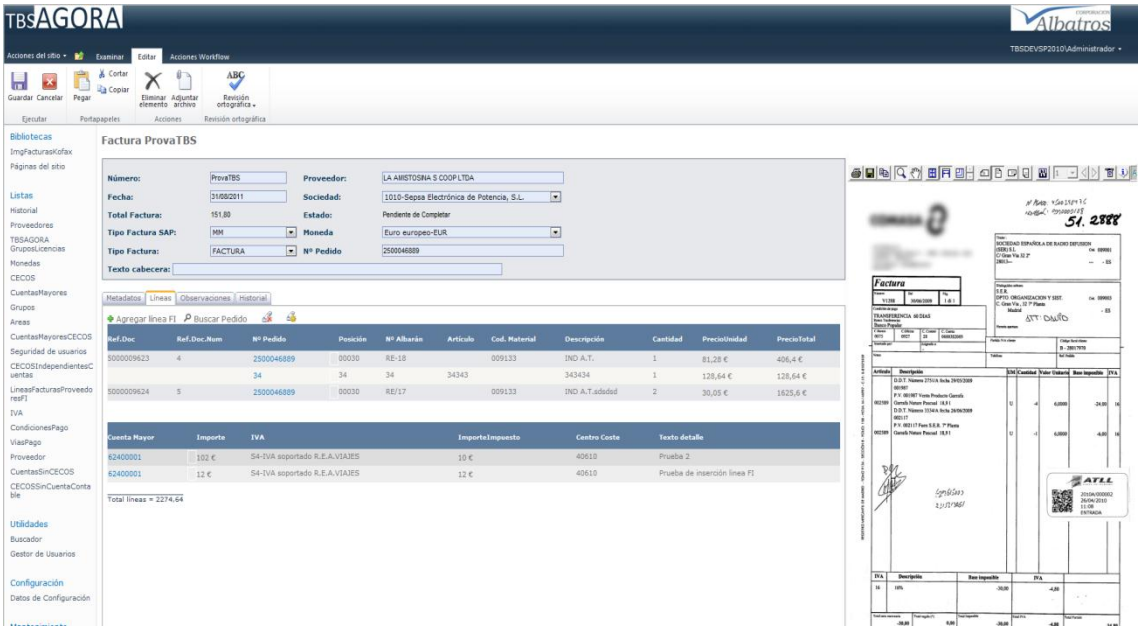


Figura 45. Pantalla de visualització de les línies d'una factura a TBS Agora

Si l'usuari creu que les dades són correctes podrà acceptar la factura mitjançant el botó "Aceptar" o refusar-la en cas contrari, amb el botó "Rechazar". Si l'aplicació detecta alguna irregularitat en la comprovació de les dades avisarà a l'usuari i li donarà

la opció d'enviar el document a revisió. Si aquest és el cas, serà un altre usuari el que pugui visualitzar la factura, l'usuari revisor.

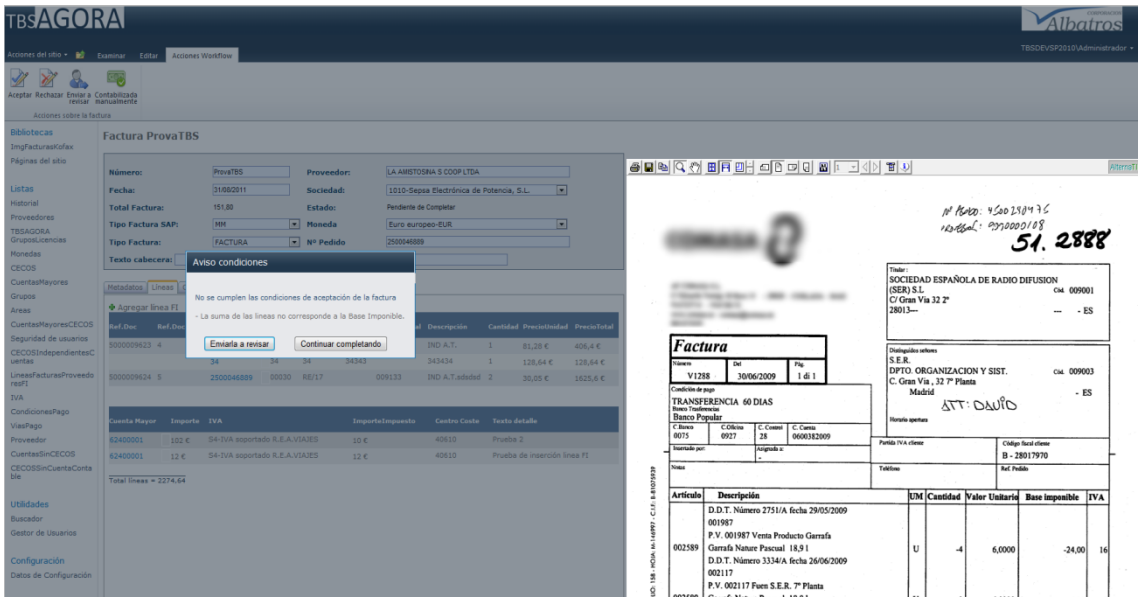


Figura 46. Avís de l'aplicació al intentar acceptar una factura amb errors

L'usuari revisor assignat a la factura podrà estar definit prèviament obtenint el responsable de la comanda o triat d'una llista d'usuaris.

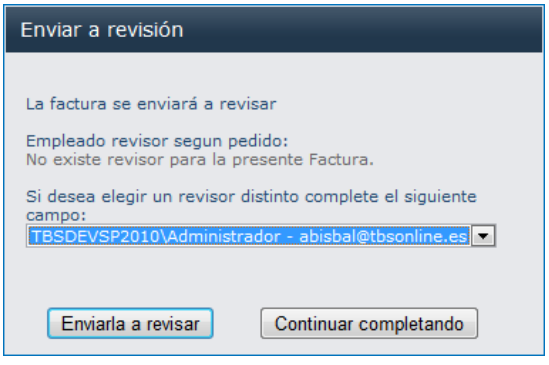


Figura 47. Finestra on es tria l'usuari revisor

Aquest usuari només podrà inserir comentaris que es posaran a la llista d'observacions associades al document, però no podrà modificar les dades. Si accepta la factura,

aquesta tornarà a ser enviada al usuari aprovador i si la refusa anirà a la llista de factures refusades. Quan una factura és canviada d'estat i passa d'un usuari a un altre s'envia automàticament un correu electrònic a l'usuari en qüestió notificant l'acció que s'ha dut a terme.

Un cop aprovada la factura per l'usuari aprovador, aquesta serà comptabilitzada a SAP, l'ERP on s'executaran els pagaments pertinents.

10. Valoració econòmica del projecte

TBS AGORA	OBSERVACIONS	Llicència	Mant. Anual de Llicència	TOTAL
TBS AGORA x SHAREPOINT - Mòdul central per 20 usuaris	Llicència base de TBS AGORA per 20 usuaris	2.490,00 €	448,20 €	2.938,20 €
TBS AGORA FACPROV	Mòdul per arxivar i gestionar Factures de Proveïdors	1.490,00 €	268,20 €	1.758,20 €
TBS AGORA eWORKFLOW X SHAREPOINT	Mòdul per a la gestió avançada de Workflows	2.990,00 €	686,25 €	3.676,25 €
TBS AGORA eREPORTS	Mòdul per a la creació d'informes i indicadors a mida	1.490,00 €	268,20 €	1.758,20 €
TBS AGORA SAP FACPROV INSERT	Mòdul per inserir Factures de Proveïdor a SAP	890,00 €	160,20 €	1.050,20 €
TBS AGORA KOFAX CONNECTOR	Mòdul per enllaçar TBS AGORA amb Kofax	490,00 €	88,20 €	578,20 €
TBS AGORA SAP DIGCERT EXTRACT	Mòdul per extreure períodes fiscals segons digitalització certificada	890,00 €	160,20 €	1.050,20 €
TBS AGORA SAP ARCHIVE LINK EXTENSION - BASE PACK	Mòdul base per convertir el repositori en compatible Archivelink	490,00 €	88,20 €	578,20 €
TBS AGORA SAP ARCHIVELINK EXTENSION - ARCHIVED DOCS 30K/yr	Paquet de 30.000 documents arxivats amb compatibilitat Archivelink	90,00 €	16,20 €	106,20 €
SUBTOTAL TBS AGORA		**11.310,00 €**	**1.497,60 €**	**12.807,60 €**
TECNOLOGIA CAPTURA DIGITAL	**OBSERVACIONS**	**Llicència**	**Mant. Anual de Llicència**	**TOTAL**
2 Escàners professionals FUJITSU Fi6130C	Escàner professional amb Kofax VRS inclòs	2.520,00 €	504,00 €	3.024,00 €
1 Kofax Image vol 300K/yr	Mòdul base de Kofax Capture Cliente Servidor	1.750,00 €	350,00 €	2.100,00 €
3 Kofax Concurrent Station	Mòdul de Kofax per a la conexió amb escàners	7.290,00 €	1.458,00 €	8.748,00 €
1 KTM Unlimited Field Extraction 30K/yr	Mòdul de Kofax per l'extracció avançada de metadades	6.600,00 €	1.320,00 €	7.920,00 €
1 Kofax KTM Add Invoice Pack 30K/yr	Mòdul de Kofax per l'extracció avançada de línies de factura	2.310,00 €	462,00 €	2.772,00 €
SUBTOTAL TECNOLOGIA CAPTURA DIGITAL		**20.470,00 €**	**4.094,00 €**	**24.564,00 €**
TECNOLOGIA DIGITALIZACIÓ CERTIFICADA	**OBSERVACIONS**	**Llicència**	**Mant. Anual de Llicència**	**TOTAL**
Legal Snap Scan + 1 certificat electrònic	Llicència base per signar factures + certificat electrònic persona jurídica	250,00 €	80,00 €	330,00 €
1 bonificació de 20.000 signatures electròniques	Bonificació per signar 20.000 factures	750,00 €		750,00 €
SUBTOTAL TECNOLOGIA DIGITALIZACIÓ CERTIFICADA		**1.000,00 €**	**80,00 €**	**1.080,00 €**
SERVEIS PROFESSIONALS				**TOTAL**
Fase d'Arrencada + Fase d'Anàlisi i Disseny				1.580,00 €
Fase de Construcció				9.950,00 €
Fase de Desplegament				5.340,00 €
TOTAL SERVEIS PROFESSIONALS				**16.870,00 €**
TOTAL PROJECTE				**55.321,60 €**

Taula 7. Desglossament de la valoració econòmica del projecte

A la taula anterior (*Taula 7*) es pot apreciar el pressupost total del projecte desglossat amb tots els components detallats i amb l'esforç realitzat separat per feses. La suma total del projecte és de **55.321,60 €**.

11. Conclusions i valoracions finals

Dins del plantejament inicial d'aquest projecte es va detectar que el procés d'administrar les factures dins d'una empresa és molt feixuc per la majoria i a l'hora de processar aquests documents les companyies consumeixen una gran quantitat de recursos en forma de personal, temps i espai.

En el marc d'aquest PFC s'ha pogut desenvolupar una solució que automatitza la captura digital d'aquests documents de manera que intervinguin els menys actors possibles, però augmentant la visibilitat i la validesa de les dades capturades al mateix temps. D'aquesta manera s'han reduït substancialment els costos de processament de les factures, a més de facilitar la gestió i l'entrada de dades d'aquestes a l'ERP empresarial. Afegint la solució de la digitalització certificada, a més, els usuaris de l'empresa client poden anar deixant d'usar progressivament el paper per dur a terme l'administració de tota aquesta documentació, evitant així, els costos de gestió documental que aquesta comporta.

Aquest PFC s'ha desenvolupat des de l'empresa TBS com a part d'una solució a les necessitats específiques de la corporació Albatros. Aquesta solució s'ha basat en la integració en una única plataforma de l'automatització de captura digital de les factures entrants, la certificació d'aquestes factures i el gestor documental TBS Agora.

Una vegada finalitzat el projecte puc afirmar que s'han assolit tots els objectius plantejats inicialment:

- ✓ **Estudi i justificació de les tècniques de reconeixement digital i del software a utilitzar.**

  S'ha elegit el software de Kofax VRS, Kofax Capture i Kofax Transformation Modules amb l'algorisme OCR Finereader.

- ✓ **Extracció del contingut dels documents escanejats amb tècniques basades en cerques sintàctiques i regles de localització.**

  Durant el període de proves s'ha pogut comprovar que el percentatge d'èxit en l'extracció del continguts de les factures està sobre el 90%.
- ✓ **Desenvolupament de la solució per validar el format de les dades resultants de la extracció.**

  Tots els camps amb format conegut són homogenis ja hagin sigut extrets automàticament o introduïts per l'usuari.
- ✓ **Validació dels resultats en conjunt mitjançant regles de negoci i dades mestres contingudes en bases de dades.**

  Les dades referents al CIF del proveïdor i a la comanda, així com els imports són validats ja hagin sigut extrets automàticament o introduïts per l'usuari.
- ✓ **Disseny i desenvolupament del formulari de validació per l'usuari.**

  El formulari de validació que visualitza l'usuari és amigable i funcionalment adequat a les necessitats del procés.
- ✓ **Transferència de les dades a bases de dades i sistemes de negoci (Gestors documentals i ERPs).**

  Les dades de les factures s'emmagatzemen a una base de dades SQL i d'allà són introduïdes al gestor documental TBS Agora a través del mòdul d'exportació.
- ✓ **Integració de la digitalització certificada al sistema.**

  El mòdul de digitalització certificada s'integra a la cua del circuit de Kofax després que l'usuari validi el lot de factures per generar la signatura electrònica corresponent i certificar cada factura.
- ✓ **Anàlisi i valoració de resultats.**

El projecte s'ha posat en producció a l'empresa Albatros durant el mes de desembre després d'haver passat per un període de proves d'aproximadament un mes. Respecte la integració de la digitalització certificada al sistema, encara està en fase de proves i està previst posar-la en producció en breu. Tant els directius com els usuaris de la corporació Albatros han tingut una actitud positiva a l'hora de valorar la nova eina de treball, ja que reconeixen que permet una gestió de les factures molt més àgil i eficient. Tot i això, i com passa amb qualsevol nova implementació d'una manera de

treballar, admeten que ha de passar un temps per poder utilitzar el sistema de la manera més òptima possible.

Finalment, puc dir que la realització d'aquest projecte m'ha aportat coneixements molt útils relacionats amb el món de la captura digital i la gestió documental. A més m'ha permès refermar el meu coneixement en els llenguatges de programació amb què s'ha desenvolupat aquest projecte.

12. Planificació del projecte

La planificació d'aquest projecte es va plantejar en un marc empresarial i la durada desitjada inicial de la realització d'aquest era aproximadament d'un any.

12.1 Descripció de les tasques

1. Estudi i justificació de les tècniques de reconeixement digital i del software a utilitzar.
2. Realització de l'anàlisi de requisits i creació de l'arquitectura de la solució.
3. Creació i configuració de tots els camps a captar digitalment dins del software de Kofax Capture i de Kofax Transformation Modules.
4. Desenvolupament del codi de l'extracció del contingut OCR dels documents amb tècniques basades en cerques sintàctiques i regles de localització.
5. Configuració i desenvolupament del codi del formateig dels camps a extreure.
6. Desenvolupament del codi de validació de les dades mitjançant regles de negoci i connexions a bases de dades habilitades per l'empresa client.
7. Creació del servei de transferència de les dades al gestor documental TBS Agora.
8. Testeig del sistema i depuració d'errors.
9. Anàlisi i valoració dels resultats amb dades demostratives de l'optimització dels processos documentals que es produeix amb la solució que presenta el projecte.

12.2 Diagrama de Gantt de la planificació prevista

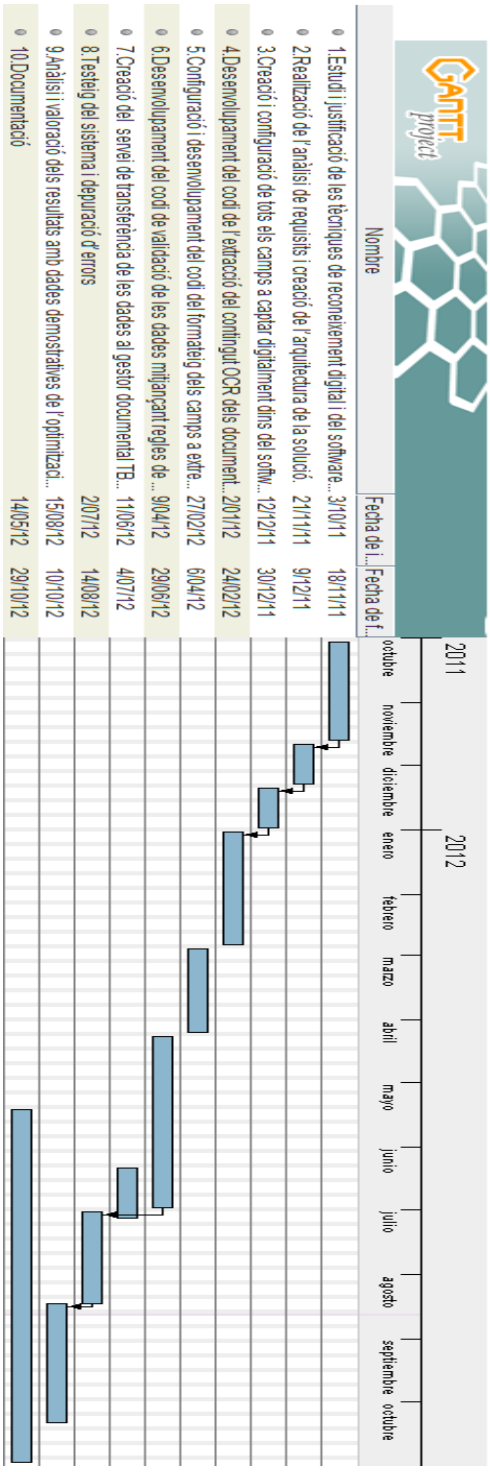


Figura 48. Diagrama de Gantt de la planificació prevista

12.3 Diagrama de Gantt de la planificació després de realitzar el projecte

S'ha afegit la tasca referent a la integració de la digitalització certificada per requeriments del projecte i perquè després de valorar els avantatges del procés integrat amb el sistema desenvolupat s'ha determinat que li dóna un valor afegit molt interessant.

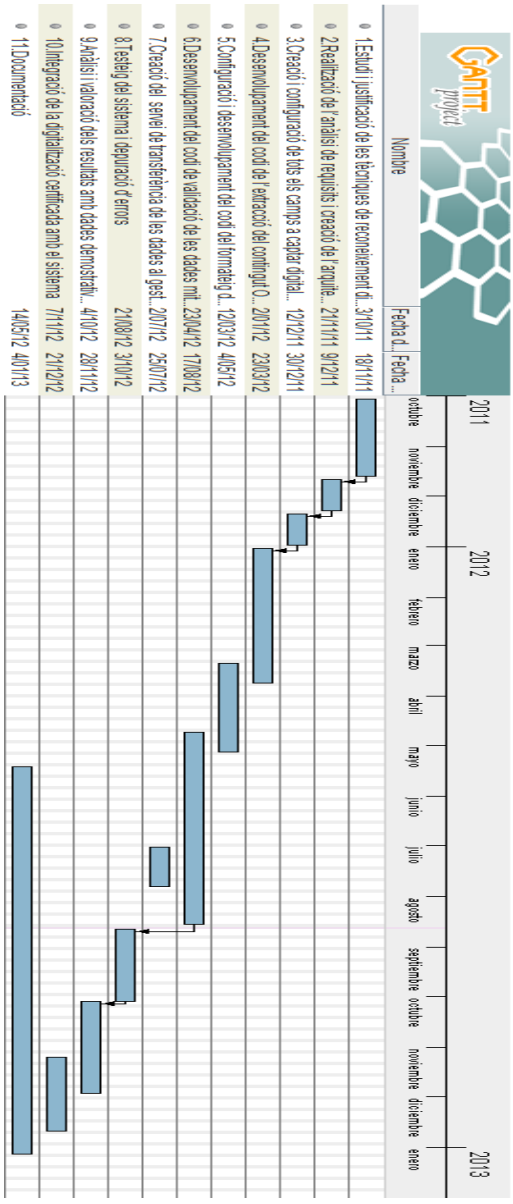


Figura 49. Diagrama de Gantt definitiu

Tenint en compte les dues planificacions es pot veure com el projecte s'ha finalitzat més tard del que s'havia previst, per tant, s'ha produït una desviació del projecte de dos mesos aproximadament. Això és degut a que les tasques de desenvolupament de l'extracció, el formateig i la validació de les dades ha requerit més temps de l'esperat i també, a que s'ha afegit una tasca nova.

13. Referències

13.1 Bibliografia

[1] Implementing Electronic Document and Record Management Systems (2007)

[2] ANF AC &TRADISE, «Integración con gestor documental» (2010).

[3] BOE número 90, Orden Ministerial EHA/962/2007 (14 de abril de 2007).

[4] SharePoint 2010 de principio a fin (2010)

13.2 Enllaços web

[1] http://es.wikipedia.org

[2] http://www.winwrap.com/web/basic/default.asp

[3] http://www.kofax.com/

[4] http://www.aberdeen.com/

[5] http://www.digitalizacioncertificada.info/

[6] http://www.aeat.es/

[7] http://www.unlibrary-nairobi.org/PDFs/KMWhitePaper_best%20practices.pdf

[i] Johannes Gensfleisch zur Laden zum Gutenberg (1398 - 1468) va ser un orfebre i inventor alemany, famós la invenció de la impremta de caràcters mòbils durant la dècada del 1450.

[ii] Informe realitzat per l'empresa Aberdeen Group: http://www.aberdeen.com/aberdeen-library/6997/RA-workflow-invoice-processing.aspx

[iii] Aberdeen Group Inc és una consultoria especialitzada en intel·ligència de negoci fundada el 1988, que tracta d'entendre les implicacions i resultats de la innovació dels processos, els avenços metodològics, les implementacions tecnològiques i la reenginyeria de negocis.

[iv] Informe comparatiu sobre la tecnologia actual en matèria de captura digital realitzada per Forrester Research, que és una companyia tecnològica independent especialitzada en la investigació de mercat que ofereix assessorament sobre l'impacte actual i potencial de la tecnologia, als seus clients i al públic.

[v] WinWrap Basic és un component .NET/WPF/COM de baix cost i de gran abast alternativa a ActiveX, VBScript, Visual Basic 6, VBA, i Sax Basic Scripting VSTA. Centenars de desenvolupadors a tot el món depenen de WinWrap Basic per proporcionar als seus clients solucions de scripting i automatització avançada.

[vi] ABBYY FineReader Engine és un potent kit de desenvolupament de programari de reconeixement òptic de caràcters que integra tecnologies d'avantguarda de reconeixement i conversió de documents de l'empresa ABBYY com: reconeixement òptic de caràcters (OCR), reconeixement intel·ligent de caràcters (ICR), reconeixement òptic de marques (OMR), reconeixement de codis de barres (OBR), representació de documents i conversió a PDF.

[vii] És un motor d'alt rendiment en reconeixement òptic, amb el que Open Text Document Technologies ha combinat els avantatges dels dos principals motors de reconeixement de caràcters internacionals, recolzar i PSW6120, en un sol producte. Tots dos motors de reconeixement s'han utilitzat durant anys a tot el món en nombroses aplicacions.

[viii] Mecanisme que permet demostrar que una sèrie de dades han existit i no han estat alterats des d'un instant específic en el temps a través d'una autoritat de segellat de temps que actua com a tercera part de confiança en el procés testificant l'existència de les dades electròniques en una data i hora concrets.

[ix] El PDF/A és un format de fitxer per a l'arxiu a llarg termini de documents electrònics. Està basat en la versió de Referència 1.4 de PDF d'Adobe Systems Inc i està definit per l'ISO 19005-1:2005.

[x] TIFF (Tagged Image File Format) és un format de fitxer per a imatges amb etiquetes. Això és perquè els fitxers TIFF contenen, a més de les dades de la imatge pròpiament dita, "etiquetes" en les quals s'arxiva informació sobre les característiques de la imatge, que serveixen per al seu tractament posterior.

[xi] JPEG (Joint Photographic Experts Group) és un estàndard de compressió i codificació d'arxius d'imatges fixes. A més de ser un mètode de compressió, és sovint considerat com un format d'arxiu, amb l'extensió Jpg.

[xii] PDF (acrònim de l'anglès portable document format, format de document portàtil) és un format d'emmagatzematge de documents, desenvolupat per l'empresa Adobe Systems. Aquest format és de tipus compost (imatge vectorial, mapa de bits i text).

[xiii] PNG (Portable Network Graphics) és un format gràfic basat en un algoritme de compressió sense pèrdua per bitmaps no subjecte a patents. Aquest format va ser desenvolupat en bona part per solucionar les deficiències del format GIF i permet emmagatzemar imatges amb una major profunditat de contrast i altres dades importants.

[xiv] Els punts per polsada (ppp), en anglès dots per inch (DPI), és una unitat de mesura per a resolucions d'impressió, concretament, el número de punts individuals de tinta que una impressora o tòner pot produir en un espai lineal d'una polsada.

[xv] Ordre Ministerial EHA/962/2007, del 10 d'abril, per la qual es desenvolupen determinades disposicions sobre facturació telemàtica i conservació electrònica de factures, contingudes en el Reial Decret 1496/2003, del 28 de novembre, pel qual s'aprova el reglament pel qual es regulen les obligacions de facturació. (BOE, 14 d'abril de 2007)